A PSYCHOPHYSICAL APPROACH TO OBJECT RECOGNITION AND ITS APPLICATION IN AIRPORT SECURITY

presented to the Faculty of Arts
of
the University of Zurich
for the degree of Doctor of Philosophy

by

Franziska Hofer
of
Bannwil / BE

Accepted on the recommendation of
Prof. Dr. Wolfgang Marx and Prof. Dr. Fred Mast

Zentralstelle der Studentenschaft
Zürich 2006

CONTENT

1 SUMMARY

This thesis is divided into two main sections: PART I "FACE AND OBJECT RECOGNITION" contains three studies, in which psychophysical experiments were conducted in the field of basic research on face and object recognition. Humans are very sensitive in detecting configural alterations in upright faces. Based on this high sensitivity in detecting configural alterations in faces, the perception of configural information was investigated in the first two studies because it could be assumed that humans have veridical percepts of configural facial information (e.g. the eye-mouth distance, inter-eye distance). Interestingly, this is not the case at all. The eye-mouth distance was overestimated up to 41 percent in the first study and 34 percent in the second study, whereas the inter-eye distance was overestimated from 16 percent (Exp. 1) to 18 percent (Exp. 2.). Furthermore, similar overestimations for upright and inverted faces were found, which contrasts with the often reported strong inversion effect for the processing of configurations in face *recognition* tasks. The results of the second study suggest an important role of well-known illusions for the large overestimations of configural information in faces.

In real life, objects are never seen in absolute isolation but are always embedded in a context. In addition, scenes can evoke specific expectations about which objects could be encountered. The third study of this thesis focused on such top-down influences in object recognition by investigating the influence of co-occurrence of episodically related objects on the identification of subsequent stimuli in non-canonical views. The main finding of this study was a decreased viewpoint-dependency for objects which were preceded by episodically related objects. Different top-down models are discussed, of which all can explain this decreased viewpoint-dependency.

The second part of the thesis PART II "HUMAN FACTORS IN AVIATION SECURITY" comprises five studies, in which different aspects of airport security are dealt with. All these studies focus on human factors in aviation security. The main value and novelty of these studies lies within the scientific visual cognition approach in airport security research. Only little psychophysical research has been done so far in this field and in this doctoral thesis it is emphasized that the use of psychophysical methods is very valuable. The knowledge of basic object recognition theories here is applied in all airport security studies to investigate different aspects of x-ray screening of passenger bags. Discriminating dangerous objects among harmless objects in passenger bags is a typical detection task, in which the forbidden object constitutes the signal and all harmless objects in a bag the noise. Therefore,

the use of psychophysical measures is appropriate to measure detection performance. In this context a conjoint objective of all the studies of the second section was to identify valid, reliable and objective measures for estimating x-ray screener detection performance from a theoretical and methodological point of view.

In the first study of this applied part, empirical results show that in x-ray screening tasks, image-based factors (namely viewpoint and superposition of the threat objects and the complexity of bags) influence x-ray screening performance and should be distinguished from knowledge-based factors. The second study is about test-psychological performance indices (validity and reliability) of the Object Recognition Test (ORT), a computer-based test developed in the framework of this thesis to measure knowledge-based factors. The third and fourth study deals with reliable and valid measures in the x-ray screening task, whereas in the former, ROC analyses suggest that nonparametric measures are more valid to measure detection performance in x-ray screening, at least for bomb detection. The latter shows that whenever using Threat Image Projection (TIP) data, data have to be aggregated over several months and a large TIP library has to be used. The last study investigated the efficiency and effectiveness of computer-based training (CBT) in x-ray screening.

2 INTRODUCTION AND OUTLINE

The architecture of the human visual system is highly adaptive to the requirements of the environment. It allows very fast and reliable automatic recognition of objects under different lighting conditions or from different distances. The ability to reliably recognize objects allows adequate reactions to the demands of our environment. Though recognizing objects is an unconscious process, the underlying mechanisms are not simple. Recognition can be described as a process, in which an internal visual memory representation has to match the internal stimulus representation (see Figure 1). If this matching process leads to an activation which exceeds the internal threshold, the object is recognized.

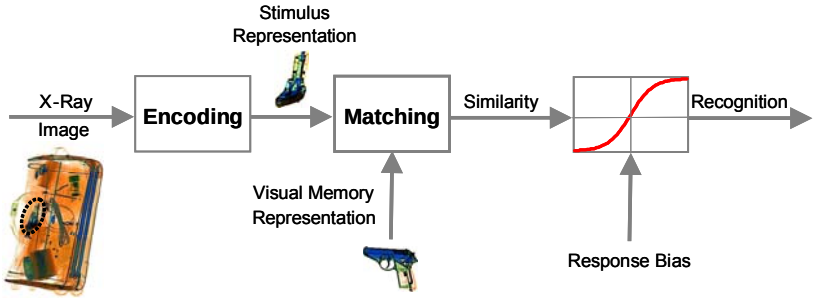

Figure 1. Illustration of the recognition process, in which an internal stimulus representation has to be matched to a visual memory representation.

Therefore, recognizing an object implies that a visual memory representation exists. All studies in this thesis emanate from such a recognition model. According to Kosslyn (1994) the process of recognition has to be distinguished from the process of identification; the first is assumed to occur based on visual memory representations whereas the latter is based on the associative memory. Therefore, identification of an object is not only a visual matching process, but comprises more elaborated mechanisms like e.g. knowing the name or function of the object. It is assumed that recognition mainly affects visual memory, whereas identification of an object covers rather multimodal information.

2.1 PART I "FACE AND OBJECT RECOGNITION"

In order to recognize an object, first of all the object has to be perceived. This sounds rather trivial at first, but an interesting question in this context is whether the perception of the object has to be veridical to guarantee a reliable recognition. This issue is covered in the first study of PART I "FACE AND OBJECT RECOGNITION", in which the perception of configural information in upright and inverted faces is investigated.

Almost forty years ago, Yin (1969) described the today very well known face inversion effect. He found that the recognition of a face is disproportionately affected by inversion when compared to the recognition of other mono-oriented objects (e.g. airplanes, houses, and stick figures of

men in motion). Current literature on face recognition research highly agrees on the fact that processing of configural information (e.g. eye-mouth distance, inter-eye distance) is strongly impaired when faces are turned upside-down. For this reason and because of other experimental results many researchers have devoted a special role to processing configural information in face *recognition* (e.g. Leder & Bruce, 2000; Murray, Yong, & Rhodes, 2000; Schwaninger, Lobmaier, & Collishaw, (2002); Schwaninger, Carbon, & Leder, (2003); Searcy & Bartlett, 1996; Sergent, 1984).

Whereas many previous studies have investigated the role of configural information for recognizing faces, the first study in PART I "FACE AND OBJECT RECOGNITION" examined the *perception* of configural information in upright and rotated faces using the method of adjustment in two experiments. Main interest of this study was to investigate if configural information in faces is perceived veridically. Humans are very sensitive in detecting configural alterations in faces (Bruce et al., 1991; Haig, 1984; Hosie et al., 1988; Kemp et al., 1990). Therefore, it could be assumed that the perception of configural information in faces is very exact. Interestingly, this was not the case at all. The eye-mouth distance was overestimated up to 41 percent and the inter-eye distance up to 16 percent. Furthermore, we found similar overestimations for upright and inverted faces. This contrasts with the often reported strong inversion effect for the processing of configurations (configural information) in face *recognition* tasks. Moreover, a comparison between upright and faces rotated 90° in the fronto-parallel plane showed that the horizontal-vertical illusion (Fick, 1851) affects the perception of the eye-mouth and the inter-eye distance less than it is the case for lines of the same length without a facial context and thus fails to provide a simple explanation for the large overestimations.

The objective of the second study in PART I "FACE AND OBJECT RECOGNITION" is on the one hand to replicate the large overestimations of perceived distances in faces found in the first study. On the other hand the effect of different facial features (eyes, nose, mouth) on the perception of distances in faces is investigated. There are different possibilities how facial features could influence the perception of distances in faces. First of all, the eyebrows and mouth might induce a Mueller-Lyer like illusion (Mueller-Lyer, 1889). This illusion has been one of the most popular visual illusions since its discovery. The perceived length of the eye-mouth distance could be influenced by the induced Mueller-Lyer like illusion. This could also be true for the inter-eye distance. The inner contours of the eyes could also induce a Mueller-Lyer illusion. The results show that the eye-mouth distance as well as the inter-eye distance is perceived greater when features which induce a

Mueller-Lyer like illusion (eyebrows and mouth for the eye-mouth distance, eyes for the inter-eye distance) are added to the head context compared to when these features are missing. Another possibility why these overestimations could arise is that the facial context influences the perceived distance between the facial features. For example Künnapas (1955) showed on the one hand that the frame size can influence perceived line length and on the other hand that different shaped surrounding fields (frames) differently affect the appearance of line length (Künnapas, 1957a; 1957b). The results suggest that there is indeed an influence of the frame of reference provided by the face context. A further possibility to consider is that the nose could induce an Oppel-Kundt like illusion (Oppel, 1855; Kundt, 1863). This illusion predicts that the length of a divided line appears longer than an undivided line of the same physical length. This hypothesis can be rejected because adding the nose onto the head context *decreased* the perceived eye-mouth distance. In addition to the role of the features on perceived distances in faces, the influence of surface based information was investigated in this study as well. Similar results could be found for line drawings as for the photographic stimuli.

In real life, objects are never seen in absolute isolation but are always embedded in a context. In addition, scenes can evoke specific expectations about which objects could be encountered. For example, viewing a scene depicting a beach and the ocean may activate different expectations which objects could be present compared to when viewing a scene depicting a forest and mountains. Therefore, we always have expectations of the objects that are more or less likely to be present in our daily life. Our knowledge influences what we expect to see. In the third study of PART I "FACE AND OBJECT RECOGNITION" such top-down influence on object recognition was investigated. Main interest of this study was the question if episodic relatedness could drive some kind of top-down priming. It is well known that object recognition is highly viewpoint-dependent (Bülthoff & Edelman, 1992; Riesenhuber & Poggio, 1999; Ullmann, 1996; Tarr & Bülthoff, 1995, 1998; Schwaninger, 2004; Wallraven, Schwaninger, Schumacher & Bülthoff; 2002). Liter and Bülthoff (1997) conducted a study in which they could show that this viewpoint-dependency is reduced in a name verification task compared to a naming task. They argue that presenting the word in the name verification task pre-activates a multiple-views representation of that object and that this pre-activation facilitates the recognition of non-canonical views. The objective of the third study of PART I "FACE AND OBJECT RECOGNITION" was to investigate if the viewpoint-dependency is also reduced when presenting a preceding object, which often co-occurs together with the target object in

reality. The hypothesis is that showing a preceding object, episodically related to the target object, also pre-activates the representation of the target object in different viewpoints. Therefore, if an episodically preceding object is shown prior to the target object, it should be easier to name the target object shown in difficult viewpoints compared to when the preceding object is not related. Exactly this was confirmed in this study. Although a viewpoint effect is still found for naming, the viewpoint-dependency is significantly reduced when an episodically related object is presented prior to the target object as opposed to a non-related one. Furthermore, a general advantage for the naming was found for the former trials. In a second experiment using a non-linguistic, contextual association task these findings could be replicated. This study shows that episodic relatedness of objects facilitates object recognition especially in non-canonical views. Therefore, the role of top-down processes in object recognition should not be underestimated.

2.2 PART II "HUMAN FACTORS IN AVIATION SECURITY"

The second part of this thesis is PART II "HUMAN FACTORS IN AVIATION SECURITY". Using psychophysical and test psychological expertise, as well as theoretical knowledge derived from object recognition theories, different aspects of human factors in aviation security are covered in six different studies.

Technological progress was enormous in the last two decades, so that today state-of-the-art technology is highly sophisticated. Current x-ray screening machines have elaborated image enhancement functions, e.g. zooming or different filtering options. Despite this highly elaborated technology a human being is still the last decision instance in the baggage control process at airports. Most computer vision scientists would agree that it will not be possible for a long time to come to adequately simulate the functions of the human brain architecture. Therefore, the use of technology is only as useful as the screener who operates it. Thus, the focus of human factors in airport security lies on the screeners' x-ray interpretation competency.

The task of a screener at an airport is to reliably detect forbidden objects among harmless objects in passenger bags. It is proposed that this visually highly demanding detection task is not only dependent on visual experience and training, but also on visual abilities such as mental rotation, visual search or figure-ground-segregation. This distinction between knowledge-based and more image-based factors is made in the first two studies in PART II "HUMAN FACTORS IN AVIATION SECURITY". Based on this distinction, the Object Recognition Test (ORT) and Prohibited Items Test (PIT) were

developed, with which x-ray image interpretation competency of novices and experts can be measured in a standardized and reliable way. In the ORT, which covers the image-based factors, only guns and knives were used because it is assumed that everyone knows the shape of guns and knives and thus, knowledge-based factors are kept constant. The ORT contains the following three image-based factors: viewpoint of the threat object, superposition of the threat object and complexity of the bag containing the threat object. The results show that novices perform only slightly poorer than experts. In addition, all three image-based factors influence detection performance of novices and experts in a similar way. Therefore, it is assumed that these factors are more strongly related to general visual abilities than to visual expertise. However, expertise might increase the required visual abilities in order to be able to cope with image difficulty resulting from effects of viewpoint, superposition and bag complexity. Therefore, taking into account possible interactions between image-based factors and expertise is also critical. Although some interactions between image-based effects and expertise were found, the effect sizes were very small compared to the main effects. In addition, large inter-individual differences were found in the ORT for novices and for experts. Furthermore, reliability coefficients (Cronbach's Alpha, Guttman Split-half) were very high. Based on these findings, this test could be a useful tool both for competency assessment of screeners as well as for pre-employment assessment purposes.

In order to have a representative sample of all threat objects in the PIT, different threat objects (according to the international prohibited items lists ICAO, ECAC, and EU) from several threat categories were used. All three image-based factors were kept constant in this test, so that mainly expertise and visual knowledge was measured. The results confirm this: compared to the ORT, very large differences in detection performance between novices and experts were obtained. The strongly improved performance of the experts confirms the assumption that this test measures knowledge and expertise rather than general visual abilities. Again, excellent reliability coefficients were found for the PIT. Therefore, this test could provide a useful tool for certification, competency and risk assessment as well as for quality control in general.

In summary, the results of these studies confirm that x-ray detection performance relies on visual abilities necessary for coping with image-based effects such as view, bag complexity and superposition. Visual experience and training are necessary to know which items are prohibited and what they look like in x-ray images of passenger bags. Both aspects are prerequisites for a good screener and can be evaluated using the ORT and PIT.

The third chapter of PART II "HUMAN FACTORS IN AVIATION SECURITY" deals with the question which psychophysical detection measure is the most fair one for the assessment of individual screener x-ray detection performance. Reliable, valid and objective measures are critical for assessing individual screener x-ray detection performance.

Detecting threat objects in passenger bags is a typical detection task, in which the threat objects constitute the signal and the harmless objects constitute the noise. A correctly identified threat object corresponds to a hit, whereas a bag, which contains no threat item judged as being harmless, represents a correct rejection. Rating a harmless bag as dangerous is a false alarm, whereas missing a forbidden object in a bag represents a miss. Only considering the number of correctly identified threat objects (hits) could lead to a biased estimation of performance due to the following reason: A high hit rate can be achieved by simply judging all bags as being dangerous. If a screener shows this response tendency, his false alarm rate (rating a bag dangerous although it does not contain a threat object) is very high. In this case security is achieved at the expense of efficiency, which would be reflected in long waiting lines at the luggage checkpoint. It would be much more beneficial to achieve a high degree of security without sacrificing efficiency. This implies a high hit rate and a low false alarm rate. Therefore, when measuring sensitivity of a screener, it is important to always look at the hit *and* false alarm rate. From a regulator's point of view, the hit rate alone is sometimes the preferred measure, because a high hit rate is indispensable for security. But from a researcher's point of view, the false alarm rate is also important, because the ratio between hit and false alarm rate reveals important information about possible response biases of the screener. Psychophysical detection theories provide several measures that take the hit and false alarm rate into account in order to achieve more valid measures of detection performance than with the hit rate alone. The use of parametric signal detection measures is still very common (e.g. McCarley et al., 2004; Swets, 1996; Schwaninger & Hofer, 2004), but several studies have used "nonparametric" A' because its computation does not require a priori assumptions of the underlying signal-noise and noise distributions (e.g. Fisk & Schneider, 1981; Prkachin, 2003; Schwaninger et al., 2004).

Receiver Operating Characteristics (ROC) can be used to test whether the assumptions of signal detection theory (SDT) are fulfilled (for detailed information on using ROCs see e.g. Gescheider, 1998; Green & Swets, 1966; MacMillan & Creelman, 1991). ROCs are produced by plotting hit rates of different criterions as a function of false alarm rates. Looking at the shape of ROCs reveals valuable information about the underlying signal-noise and

noise distributions and therefore delivers useful information about which psychophysical measure should be used. X-ray screening data coming from computer based training with improvised explosive devices (IEDs) were analyzed prior to and after training. The following detection measures were compared in this study: pHit, d', Δm, Az, A', $p(c)_{max}$. Interestingly, non-standardized ROC curves could be fitted very well by two straight lines, just as would be predicted by the two-state low threshold theory of Luce (1963). In addition, standardized ROCs deviated from linearity both before and after training of IED detection. These results challenge the validity of SDT measures as estimates of threat detection performance, at least what concerns detection of IEDs in x-ray images of passenger bags. But it has to be noted that all five psychophysical measures compared in this study were strongly correlated in most cases. A' and d' correlated with $r >= .75$ in all four test conditions. And in the first study in PART II "HUMAN FACTORS IN AVIATION SECURITY" even higher correlations between A' and d' were found ($r > .90$). It certainly remains to be investigated whether these results can be replicated with different stimulus material and other threat items. In any case, our findings suggest that additionally to SDT measures other detection performance measures should be considered. As mentioned above, the calculation of A' requires no a priori assumptions about the underlying distributions, which has often been regarded as an advantage over SDT measures such as d' and Δm.

As already mentioned, technology has evolved remarkably during the last two decades. One relatively new technology is threat image projection (TIP). This is a software function of state-of-the art x-ray machines that allows on the job measurement of x-ray detection performance.

If a screener detects the fictional threat item within a certain time, the answer is considered a hit, whereas missing a TIP-image is considered a miss. Non-Tip alarms are registered in cabin baggage screening (CBS) if a screener gives a "threat present" response when no TIP image was shown. In some hold baggage screening (HBS) systems not only threat x-ray images but also non-threat x-ray images of passenger bags are shown. In this case, false alarms as well as correctly rating bags harmless are additionally recorded in TIP report files.

TIP data could be a very valuable source for quality control, risk analysis and assessment of individual screener performance. Especially for the latter purpose, reliability of measurement is of special importance. This was examined in the fourth study analyzing reliability coefficients of CBS and HBS data. In this study, all reliability analyses were done only with the hit rate

due to the following reasons: First, if a screener detects a real threat in a bag when no TIP image was shown, this is recorded as a non-TIP alarm in the TIP report. In this case the response should count as a (true) hit and certainly not as a false alarm. Therefore, it is not possible to get a valid measure of the false alarm rate from CBS TIP reports because the individual non-TIP alarm rate does not completely match the individual false alarm rate. Second, because correctly judging a bag to be harmless is not written into the CBS TIP report, the individual non-TIP alarm rate has to be estimated based on the averaged TIP to bag ratio, which can further reduce the internal validity of the estimates. As mentioned above, looking only at the hit rate as the sole basis for assessing x-ray detection performance can lead to biased estimates. The non-TIP alarm rates (CBS) and false alarm rates (HBS) are also reported to illustrate differences between individuals in their response tendencies.

For CBS data very low reliability values (all r <=.58) were found, even for data aggregated over seven months. It is important to note that the hit rate was very high and only small inter-individual variance was observed. In contrast to CBS, the reliabilities for HBS data were very high (for data aggregated over 6 or more months, the correlation was > .90). Compared to the CBS data, the hit rate of the HBS TIP data was not at ceiling, and larger standard deviations could be observed.

Several possible reasons why the CBS data compared to HBS has such low reliability are discussed in this chapter. This study also showed that there are substantial inter-individual differences in non-TIP alarm rates (CBS) and false alarm rates (HBS), which also affects the validity of hit rates as a measure of x-ray detection.

The last study of this thesis is an evaluation study to measure the effectiveness and efficiency of a computer based training system for IED detection in x-ray screening. Training and visual expertise are very important aspects for interpreting x-ray images of passenger bags. This is especially evident when IED detection is concerned, because the visual appearance of IEDs in x-ray images varies enormously. Therefore, efficient and effective visual training is essential to guarantee a high level of airport security. Using a Latin square counterbalanced design, four groups of screeners with comparable baseline IED detection performance were trained about twice a week for 20 minutes with X-Ray Tutor during six months. X-Ray Tutor is an individually adaptive computer based training, which was developed in close collaboration between visual cognition scientists and aviation security experts. In order to measure training effectiveness, four performance tests were conducted with new X-ray images during the six months. Remarkable increases in detection performance were observed. Relative increase in

detection performance as compared to the first test was 71 percent after an average of 28 training sessions during the six months period. For a subgroup of 52 screeners, who on average went through 31 training sessions, relative increase in detection performance was even higher, i.e. 84 percent. Simultaneously to this high detection performance increase reaction times for correctly identified dangerous bags decreased significantly during the training. This could not be observed for harmless bags correctly rated harmless. Therefore, the results of this study suggest that training leads to stronger visual memory representations and not to a general increase in visual processing capacity.

References

Bruce, V., Doyle, T., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition, 38,* 109-144.

Bülthoff, H. H. & Edelman, S. 1992 Psychophysical support or a two-dimensional view interpolation theory of object recognition. *Proc. Natl Acad. Sci. USA 89*, 60–64.

Fick, A. (1851). *De errore quodam optico asymenetria bulbi effecto.* Marburg: J.A. Kochii.

Fisk, A. D., & Schneider, W. (1981). Control and Automatic Processing during Tasks Requiring Sustained Attention: A New Approach to Vigilance. *Human Factors, 23,* 737-750.

Gescheider, G. A. (1998). *Psychophysics: The Fundamentals (3rd Ed),* Mahwah, NJ: Lawrence Erlbaum Associates.

Green, D. M., & Swets, A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception, 13,* 505-512.

Hosie, J.A., Ellis, H.D., & Haig, N.D. (1988). The effect of feature displacement on the perception of well-known faces. *Perception 17,* 461-474.

Kemp, R., McManus, C., & Pigott, T. (1990). Sensitivity to the displacement of facial features in negative and inverted images. *Perception, 19,* 531-543.

Kosslyn, S. M. (1994). *Image and brain. The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Kundt, A. (1863). Untersuchungen ueber Augenmass und optische Taeuschungen. *Poggendorff Analle, 120,* 118-158.

Künnapas, T. M. (1955). Influence of frame size on apparent length of a line. *Journal of Experimental Psychology, 30,* 168-170.

Künnapas, T.M. (1957a). The vertical-horizontal illusion and the visual field. *Journal of Experimental Psychology, 53(6),* 405-407.

Künnapas, T.M. (1957b). Vertical-horizontal illusion and surrounding fields. *Acta Psychologica, 13,* 35-42.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: the role of configural information in face recognition. *Quarterly Journal of Experimental Psychology Section A -Human Experimental Psychology, 53,* 513-536.

Liter, J.C. &Bülthoff, H.H. (1997). View canonicality affects naming but not name verification of common objects. *Tehnical Report No. 051,* Max-Planck Institute for Biological Cybernetics, Tuebingen.

Luce, R. D. (1963). A threshold theory for simple detection experiments, *Psychological Review, 70*, 61-79.

MacMillan, N. A., & and Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge: University Press.

McCarley, J. S., Kramer, A., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual Skills in Airport-Security Screening. *Psychological Science, 15*, 302-306.

Mueller-Lyer, F.C. (1889). Optisches Urteilstaeuschungen. Dubois-Reymonds *Archive für Anatomie und Physiologie (Suppl.),* 263-270.

Murray, J.E., Yong, E., & Rhodes, G. (2000). Revisiting the perception of upside-down faces. *Psychological Science, 11,* 492-496.

Oppel, J.J. (1855). Über geometrisch-optische Täuschungen. *Jahresbericht des Frankfurter Vereins,* 37-47.

Prkachin, G. C. (2003). The effects of orientation on detection and identification of facial expressions of emotion. *British Journal of Psychology, 94*, 45-62.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex : WIT Press.

Schwaninger, A. (2004). Objekterkennung und Signaldetektion. In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 106-130). Bern: Huber.

Schwaninger, A., Carbon, C.C., & Leder, H. (2003). Expert face processing: Specialization and constraints. In G. Schwarzer & H. Leder, *Development of face processing* (pp. 81-97) , Göttingen: Hogrefe.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38*, 258-264.

Schwaninger, A., Lobmaier, J., Collishaw, S.M. (2002). Component and configural information in face recognition. *Lectures Notes in Computer Science, 2525,* 643-650.

Searcy, J.H., & Bartlett, J.C. (1996). Inversion and processing of components and spatial relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 904-915.

Sergent, J. (1984). An investigation into component and configurational processes underlying face recognition. *British Journal of Psychology, 75,* 221-242.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics – Collected Papers*, Mahwah, NJ: Lawrence Erlbaum Associates.

Tarr, M. J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? J*ournal of Experimental Psychology: Human Perception and Performance, 21(6)*, 1494-1505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. In M.J. Tarr & H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (pp. 1-20). Cambridge, MA: MIT Press.

Ullman, S. (1996). *High-level vision. Object recognition and visual cognition.* Cambridge, MA: MIT Press.

Wallraven, C., Schwaninger, A., Schuhmacher, S. & H.H. Bülthoff. View-Based Recognition of Faces in Man and Machine: Re-visiting Inter-Extra-Ortho. *In Proc. BMCV'02,* 2002.

Yin, R.K. (1989). Looking at upside-down faces. *Journal of Experimental Psychology, 81,* 141-145.

PART I
FACE AND OBJECT RECOGNITION

3 CONFIGURAL INFORMATION IS PROCESSED DIFFERENTLY IN PERCEPTION AND RECOGNITION OF FACES

3.1 ABSTRACT

Several previous studies have stressed the importance of configural information in face recognition. In this study the perception of configural information was investigated. Large overestimations were found when the eye-mouth distance and the inter-eye distance had to be estimated. Whereas configural processing is disrupted when inverted faces have to be recognized the perceptual overestimations persisted when faces were inverted. These results suggest that processing of configural information is different in perceptual as opposed to recognition tasks.

3.2 INTRODUCTION

Processing facial information is one of the most relevant skills in everyday life. Although faces seem to look quite different from each other, they do in fact form a very homogenous stimulus class when seen from an image-based point of view. Each face has the same components (eyes, nose, mouth etc.) in the same basic arrangement. Therefore, reliably recognizing faces entails detecting subtle differences between components and their spatial interrelationship (configural information). Whereas component processing seems to be relatively unaffected by orientation changes, the processing of configural information is strongly impaired when faces are rotated. Indeed, many researchers have argued that turning faces upside-down disrupts configural processing much more than component processing (e.g. Leder & Bruce, 2000; Murray, Yong, & Rhodes, 2000; Schwaninger & Mast, 1999; Searcy & Bartlett, 1996; Sergent, 1984). More than 30 years ago, it was found that face recognition is disproportionately affected by inversion when compared to the recognition of other mono-oriented objects such as airplanes, houses, and stick figures of men in motion (Yin, 1969). Since face recognition is highly orientation-sensitive and the processing of configural information is strongly impaired when faces are turned upside-down, many researchers have devoted a special role to the processing of configural information in face recognition. Whereas many previous studies have investigated the role of configural information for recognizing faces this study examines the perception of configural information in upright and rotated faces.

3.3 EXPERIMENT 1

Face recognition is characterized by a high sensitivity for configural

information. For example Haig (1984) revealed for unfamiliar faces that configural alterations, which were induced by changing the distance between facial components, are sometimes detected at the visual acuity threshold level. Similar results were reported by Hosie, Ellis and Haig (1988) for familiar faces.

Whereas these studies were concerned with detecting alterations of configural information in faces the aim of Experiment 1 was to investigate whether human observers have a veridical percept of configural information.

3.3.1 METHOD

3.3.1.1 PARTICIPANTS

Twenty undergraduates from the University of Zurich voluntarily participated in this study. The participants were randomly assigned to two groups of 10 participants. All had normal or corrected to normal vision.

3.3.1.2 MATERIALS AND PROCEDURE

Photographs were made from 10 persons (5 female) who had agreed to be photographed and to have their pictures used in psychology experiments. The faces in the original grayscale pictures were front facing and had a neutral expression. In digital versions the hair was removed and the faces were placed on a black background.

The experiments were conducted in a dimly lit room. The viewable screen area on the TFT display was limited to a 750*750 pixel square (23.5° of visual angle) by a cardboard covering the 14.1 inch screen. The viewing distance was maintained by a head rest so that the center of the screen was at eye height of participants and the height and width of displayed faces covered 8.5° and 6.7° of visual angle, respectively.

The method of adjustment was applied. The length of a simultaneously presented white line (comparison stimulus) had to be adjusted in order to appear as long as the standard stimulus. For half of the participants the standard stimulus was the eye-mouth distance, for the other half of participants the standard stimulus was the inter-eye distance (Figure 1). The latter was defined as the distance

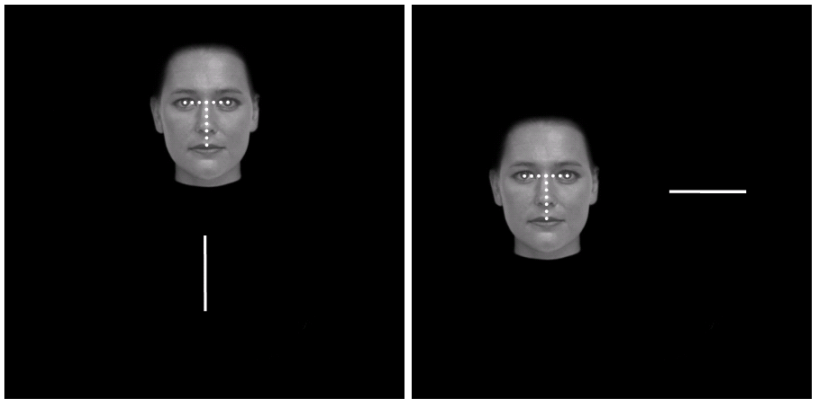

Figure 1: The two positions of standard and comparison stimuli (line) for one face as standard stimulus. Dotted lines indicate the inter-eye distance and eye-mouth distance and were not shown in the experiments.

between the pupils (mean distance was 84 pixel or 2.6° of visual angle). The eye-mouth distance was defined as the vertical distance between the point in the middle of the upper contour of the mouth and the point where a vertical line through this point would cross a horizontal line connecting the two pupils (mean distance was 86 pixel or 2.7° of visual angle). Adjustments were made with the preferred hand by turning a small wheel on a mouse device. Each trial was started by pressing a button on this device. The adjustment line (comparison stimulus) was one pixel in width and its initial length was either 20 or 180 percent of the standard stimulus. For the two standard stimuli (inter-eye distance and eye-mouth distance) the line comparison stimulus was presented horizontally to the right of the standard stimulus and vertically on bottom of the standard stimulus (Figure 1).

There were 40 trials for each standard stimulus: 10 (faces) * 2 (initial line lengths) * 2 (positions). The order of faces, initial line lengths, and line positions was counterbalanced across participants using Latin squares.

3.3.2 RESULTS AND DISCUSSION

Individual data were averaged across the two measurement conditions, the two initial line lengths and the ten faces.

The eye-mouth distance was overestimated by 39 percent (SE = 5.96) and the inter-eye distance by 11 percent (SE = 4.02)[1].

Several previous studies have found a high sensitivity for detecting subtle configural changes (Bruce, Doyle, Dench, & Burton, 1991; Haig, 1984; Hosie et al., 1988; Kemp, McManus, & Pigott, 1990). The large overestimations revealed in the present study indicate that the ability of skilled perceptual discrimination does not necessarily imply very precise veridical percepts. In contrast, the overestimations found in Experiment 1 are of a magnitude that exceeds most known perceptual size illusions (e.g. Coren & Girgus, 1978).

3.4 EXPERIMENT 2

The processing of configural information in recognition and detection tasks is strongly impaired when faces are inverted (Leder & Bruce, 2000; Rhodes, Brake, & Atkinson, 1993; Schwaninger & Mast, 1999; Sergent, 1984; Young, Hellawell, & Hay, 1987). If there was a difference in the perception of configural distances between upright and inverted faces, then the face inversion effect could be related to perceptual processes. In contrast, if the

[1] Based on the horizontal vertical illusion (HVI), the horizontal vs. vertical placement of the comparison line could be expected to influence the adjustments. Indeed, separate analyses for the two measurement conditions (horizontal vs. vertical placement of the comparison line) revealed significant effects for both facial distances: When the comparison line was oriented horizontally (as opposed to oriented vertically), the overestimation of the eye-mouth distance was 10 percent larger, $t(9) = 2.98$, $p < .05$, and the overestimation of the inter-eye distance was 8 percent larger, $t(9) = 3.71$, $p < .01$. In order to reduce such effects based on the placement of the comparison line, the data were averaged across the two measurement conditions.

overestimations found in Experiment 1 would persist to the same degree in inverted faces, the orientation-dependent nature of configural processing in face recognition can not be explained based on limitations on the perceptual level.

A second aim of Experiment 2 was to investigate a possible role of the horizontal vertical illusion (HVI). This perceptual phenomenon has been first reported by Fick (1851) and refers to the observation that vertical lines or distances appear longer than horizontal ones of the same physical length. The HVI has also been shown to affect the perception of various objects including complex stimuli such as houses (e.g. Higashiyama, 1996; Yang, Dixon, & Proffitt, 1999). In Experiment 2 a potential effect of the HVI on the perception of configural information in faces was investigated by showing the faces in four angles of clockwise rotation (0°, 90°, 180°, 270°) and comparing the overestimations of configural information to the overestimation of line length.

3.4.1 METHOD

3.4.1.1 PARTICIPANTS

Twenty-four undergraduates from the University of Zurich volunteered in this study. All had normal or corrected to normal vision.

3.4.1.2 MATERIALS AND PROCEDURE

One male and one female face from Experiment 1 served as stimuli. The experimental setup was identical to Experiment 1. The length of a simultaneously presented white line (comparison stimulus) had to be adjusted in order to appear as long as the standard stimulus. For 12 randomly selected participants the standard stimulus was the inter-eye distance and the eye-mouth distance of the simultaneously presented face (both distances were 83 pixel or 2.6° of visual angle). The distances were explained to the participants the same way as in Experiment 1. In order to ensure that the participants understood the definitions of the distances precisely, the distances were indicated with white lines on a face presented on a cardboard above the computer screen. The eye-mouth and the inter-eye distance were adjusted in separate blocks, counterbalanced across subjects. For the other 12 randomly selected participants the standard stimulus was a simultaneously presented white line that was one pixel in width and 83 pixels in length. Adjustments were made as in Experiment 1. Again, the adjustment line (comparison stimulus) was one pixel in width and its length was either 20 or 180 percent of the standard stimulus. The comparison stimulus was presented horizontally to the right or left of the standard stimulus and vertically on top

or bottom of the standard stimulus, so that in half of the trials the comparison line was at the same orientation as the facial distance, whereas in the other half of the trials the comparison line was perpendicular to it.

The standard stimuli were presented in four angles of clockwise rotation (0°, 90°, 180°, 270°) around their center.

There were two blocks of 64 trials resulting in 128 trials for the group in which the eye-mouth distance and the inter-eye distance served as standard stimuli: 2 (adjustments for the male and female face) * 2 (initial lengths of comparison stimulus) * 4 (positions of standard and comparison stimuli) * 4 (angles of rotation of the standard stimulus) * 2 (blocks: eye-mouth distance and inter-eye distance). Since for the second group the standard stimulus was a line instead of facial distances, only one block (64 trials) was used: 2 (adjustments) * 2 (initial lengths of comparison stimulus) * 4 (positions of standard and comparison stimuli) * 4 (angles of rotation of the standard stimulus). The order of positions, rotations, length of comparison stimulus as well as order of faces and blocks (group one only) was counterbalanced across participants using a mixed Latin square design.

3.4.2 RESULTS AND DISCUSSION

Individual data were averaged across the four measurement conditions, the two initial lengths of the comparison stimulus as well as the two adjustments.

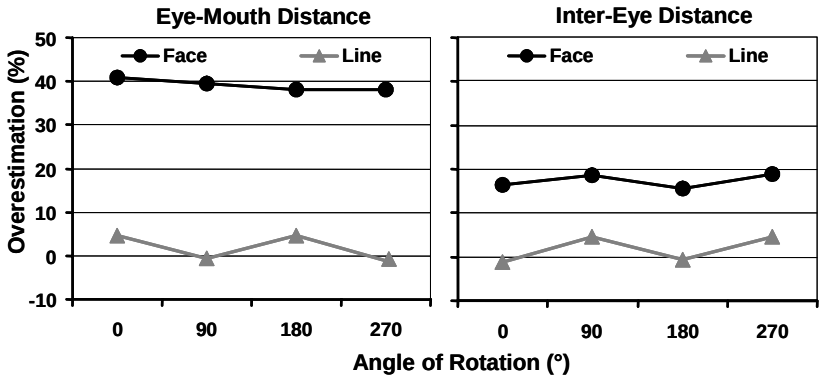

Figure 2 Large overestimation of configural information in faces and the effect of orientation. Left: eye-mouth distance, right: inter-eye distance.

As shown in Figure 2, the line was over-estimated when presented vertically and slightly underestimated when presented horizontally. This result reflects the well known horizontal vertical illusion.

The results from Experiment 1 were replicated.

The eye-mouth distance was overestimated by 41 percent and the inter-eye distance by 16 percent in upright faces[2]. A two factor analysis of variance (ANOVA) with standard stimulus (eye-mouth distance vs. line) as between-subjects factor and orientation as within-subjects factor revealed that the eye-mouth distance was much more overestimated than the line, $F(1, 22) = 13.79$, $MSE = 2422.01$, $p < .01$. There was also a main effect of orientation[3], $F(2.33, 51.23) = 18.89$, $MSE = 6.16$, $p < .001$, and an interaction between orientation and standard stimulus (eye-mouth distance vs. line), $F(2.33, 51.23) = 10.73$, $p < .001$. As indicated by the interaction the HVI affected perceived line length more than the perception of the eye-mouth distance.

A separate two factor analysis of variance (ANOVA) with standard stimulus (inter-eye distance vs. line) as between-subjects factor and orientation as within-subjects factor revealed larger overestimations of the inter-eye distance than of line length, $F(1, 22) = 4.86$, $MSE = 1177.18$, $p < .05$. There was a main effect of orientation, $F(2.28, 50.09) = 26.90$, $MSE = 6.63$, $p < .001$. Again, there was an interaction between orientation and standard stimulus (inter-eye distance vs. line), $F(2.28, 50.09) = 3.19$, $p < .05$, confirming that also the perception of the inter-eye distance is less affected by the HVI than the perception of lines.

The effects of orientation were further examined using Bonferroni corrected pairwise comparisons of means (Table 1).

		Eye-Mouth Distance			Inter-Eye Distance		
(I) ANGLE	(J) ANGLE	MD (I-J)	SE	p	MD (I-J)	SE	p
0	90	1.623	1.306	1.000	-2.170	1.108	.456
0	180	2.843	1.110	.159	0.855	1.117	1.000
0	270	2.930	1.046	.103	-2.496	1.044	.215
90	180	1.220	1.045	1.000	3.025	1.213	.179
90	270	1.306	0.708	.552	-0.326	0.623	1.000
180	270	0.087	0.776	1.000	-3.351	1.139	.080

Table 1 Bonferroni corrected pairwise comparisons between the four angles used in Experiment 2. *Note.* MD = mean difference, SE = standard error.

There were no significant differences, neither for the inter-eye distance nor for the eye-mouth distance. More specifically, the large overestimations were similar for upright and inverted faces[4], which contrasts with the often

[2] As mentioned in footnote 1, the placement of the comparison line had a modulatory effect on the overestimations in Experiment 1. Similar effects were found in Experiment 2. On average, the overestimation was 8 percent larger for horizontal vs. vertical placements of the comparison line. This effect was comparable across conditions since separate ANOVAs for the eye-mouth and the inter-eye distance with measurement condition as within-subjects factor (horizontal vs. vertical placement of the comparison line) and standard stimulus (line vs. facial distance) as between-subjects factor gave no significant interactions between these two factors. As in Experiment 1, we averaged across the two measurement conditions in order to reduce modulatory effects caused by the placement of the comparison line.

[3] In all analyses of this study, if Mauchly's (1940) test of sphericity showed a significant deviance ($\alpha \geq .25$) from equicorrelation for a repeated factor or for a combination of factors including at least one repeated factor, Greenhouse and Geisser's (1959) Epsilon was used to adjust the degrees of freedom for the averaged tests of significance.

[4] However, the small mean difference of 2.8 percent between adjustments of the eye-mouth distance for upright vs. inverted faces was significant when a paired-samples t-test was used (without Bonferroni adjustment for multiple comparisons), $t(11) = 2.56$, $p < .05$.

reported strong inversion effect for processing configuration in face recognition tasks.

3.5 General Discussion

Many previous studies have stressed the importance and orientation-sensitivity of configural processing for recognizing faces. In the present study we investigated the *perception* of configural information in faces and found new and surprising results. While people are very sensitive in detecting configural differences (Bruce et al., 1991; Haig, 1984; Hosie et al., 1988; Kemp et al., 1990) our study shows that configural information is not perceived veridical but is instead overestimated by 11-41 percent. Inversion strongly impairs configural processing in detection and recognition tasks (e.g. Leder & Bruce, 2000; Murray et al., 2000; Rhodes et al., 1993; Schwaninger & Mast, 1999; Searcy & Bartlett, 1996; Sergent, 1984; Young et al., 1987). In contrast, our study revealed that the perception of configural information is much less orientation-sensitive. Moreover, a comparison between overestimations of distances in upright and in 90° rotated faces showed that the HVI affects the perception of the eye-mouth and the inter-eye distance less than it is the case for lines of the same length and thus fails to provide a simple explanation of the large overestimations.

In short, this study revealed a new and large perceptual illusion in faces and indicates that configural processing does not obey the same rules in perceptual tasks as opposed to detection and recognition tasks.

3.6 References

Bruce, V., Doyle, T., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition, 38,* 109-144.

Coren, S., & Girgus, J.S. (1978). *Seeing is Deceiving: The Psychology of Visual Illusions*. Hillsdale, NJ: Erlbaum.

Fick, A. (1851). De errore quodam optico asymenetria bulbi effecto. Marburg: J.A. Kochii.

Greenhouse, S.W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika, 32,* 95-112.

Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception, 13,* 505-512.

Higashiyama, A. (1996). Horizontal and vertical distance perception: The discorded-orientation theory. *Perception & Psychophysics, 58,* 259-270.

Hosie, J.A., Ellis, H.D., & Haig, N.D. (1988). The effect of feature displacement on the perception of well-known faces. *Perception 17,* 461-474.

Kemp, R., McManus, C., & Pigott, T. (1990). Sensitivity to the displacement of facial features in negative and inverted images. *Perception, 19,* 531-543.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: the role of configural information in face recognition. *Quarterly Journal of Experimental Psychology Section A -Human Experimental Psychology, 53,* 513-536.

Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics, 11,* 204-210.

Murray, J.E., Yong, E., & Rhodes, G. (2000). Revisiting the perception of upside-down faces. *Psychological Science, 11,* 492-496.

Rhodes, G., Brake, S., & Atkinson, A.P. (1993). What's lost in inverted faces? *Cognition, 47,* 25-57.
Schwaninger, A., & Mast, F. (1999). Why is face recognition so orientation-sensitive? Psychophysical evidence for an integrative model. *Perception, 28* (Suppl.), 116.
Searcy, J.H., & Bartlett, J.C. (1996). Inversion and processing of components and spatial relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 904-915.
Sergent, J. (1984). An investigation into component and configurational processes underlying face recognition. British Journal of Psychology, 75, 221-242.
Yang, T.L., Dixon, M.W., & Proffitt, D.R. (1999). Seeing big things: Overestimation of heights is greater for real objects than for objects in pictures. Perception, 28, 445-467.
Young, A.W., Hellawell, D., & Hay, D.C. (1987). Configural information in face perception. Perception, 16, 747-759.
Yin, R.K. (1969). Looking at upside-down faces. Journal of Experimental Psychology, 81, 141-145.

4 WHY CONFIGURAL INFORMATION IN FACES IS OVERESTIMATED BY 15-40%

4.1 ABSTRACT

Several previous studies have annotated the importance of configural processing for face *recognition*. In an earlier study (Schwaninger, Ryf, & Hofer, 2003) we showed that the *perception* of the configural information in faces (eye-mouth and the inter-eye distance) is overestimated to a large extend. We replicated these results and investigated to what extend the features of a face contribute to this large illusion and whether the same effects are found when a line drawing instead of a realistic photograph is used. The results suggest an important role of well-known illusions for the large overestimation of configural information in faces.

4.2 INTRODUCTION

Many previous studies have investigated the role of configural processing for recognizing faces (e.g, Leder & Bruce, 2000; Murray, Yong, & Rhodes, 2000; Schwaninger & Mast, 1999; Searcy & Bartlett, 1996; Sergent, 1984). All these studies emphasize that the processing of configural information, like the inter-eye or the eye-mouth distance, is more affected by rotation (inversion) than the processing of feature information (i.e. the parts of a face like the eyes, nose and mouth). This impaired processing of configural information is considered to be the most important factor for the diminished recognition of inverted faces compared to other mono-oriented objects turned upside-down.

But what about our perception of configural information? In a previous study we could show that we do not have veridical percepts of the eye-mouth distance or inter-eye distance in faces, but overestimate this configural information by 11-41 percent (Schwaninger, Ryf, & Hofer, 2003). Interestingly, compared to the clear inversion effect for face recognition (Yin, 1969), we only found a very small effect of inversion for the perception of configural information in faces. A further result is that an explanation based on the horizontal-vertical illusion (HVI) first described by Fick (1851) cannot be solely accountable for the large overestimations.

One aim of the present study was to replicate the findings of the study mentioned above, whereas the main focus was to investigate the role of different features (eyes, nose, mouth) onto the perception of distances in faces. There exist different conceivable possibilities how facial features could influence the perceived distances in faces and result in the large overestimation of distances in faces. For example the eyebrows and mouth

might induce a Mueller-Lyer like illusion (dotted lines in Figure 2d and 3d, not shown in the experiments).

It is also possible that besides the eye-mouth distance, the perceived inter-eye distance is influenced by the contours of the eyes inducing also a Mueller-Lyer like illusion (dotted lines in Figure 2c and 3c, not shown in the experiment). A different reason why distances in faces are not perceived veridical, but are highly overestimated could be related to the finding that the context or frame of reference (e.g. different shaped surroundings) can alter the perceived length of a line. Künnapas (1955) showed not only that the frame size can influence perceived line length but also that different shaped surrounding fields (frames) affect the appearance of line length differently (Künnapas, 1957a,b). Therefore, it is possible that the head context provides a frame of reference that alters the perceived distances in faces. A third possibility to consider is that the nose could induce an Oppel-Kundt like illusion (Oppel, 1855; Kundt, 1863). This illusion attests that the length of a line, which is divided, appears longer than an undivided line of the same physical length. Since the nose divides the distance between the eyes and the mouth, it is conceivable that the nose induces an elongated appearance of this distance. A further thinkable explanation for an elongated percept of the eye-mouth distance in a face is that the nose conciliates three dimensional depth information, which might change the perceived eye-mouth distance, too.

Besides the interest to examine the influence of feature information onto perceived distances in faces, an additional objective of Experiment 2 was to examine a possible role of texture information. This was tested by comparing the perceived lengths of configural information in a line drawing of a face to a realistic photograph of a face.

In both experiments we used the psychophysical method of adjustment to test these possible influences on the perception of configural information in faces.

4.3 EXPERIMENT 1

Experiment 1 served as replication of the previously found result that distances in faces (eye-mouth distance, inter-eye distance) are overestimated by an amount which exceeds most known perceptual size illusions (Schwaninger, Ryf, & Hofer, 2003).

4.3.1 METHOD

4.3.1.1 PARTICIPANTS

Twenty undergraduates from the University of Zurich volunteered in this study. The participants were randomly assigned to two groups of 10

participants each. All had normal or corrected to normal vision.

4.3.1.2 MATERIALS AND PROCEDURE

We used photographs of faces of 10 persons (5 females) who had agreed to be photographed and to have their pictures used in psychology experiments as stimulus material. The original grayscale pictures showed the faces from the front with a neutral expression. The hair was removed digitally and the faces were placed on a black background.

Participants were seated in front of a computer in a dimly lit room. The center of the screen was at eye height of participants. To avoid external context effects the 14.1 inch screen was limited to a 750*750 pixel square (23.5° of visual angle). A head rest fixed participants' head so that the height and width of presented faces covered approximately 8.5° and 6.7° of visual angle and that the distance to the screen maintained constant.

The task of participants consisted of estimating the length of configural information in faces. Using the method of adjustment participants had to adjust the length of a white line (comparison stimulus) in order to perceive it as long as a standard stimulus. For participant group 1 the standard stimulus was the eye-mouth distance and for participant group 2 this was the inter-eye distance.

The latter was defined as the distance between the pupils (mean distance was 84.4 pixel or approx. 2.6° of visual angle), whereas the former was defined as the vertical distance between the point in the middle of the upper contour of the mouth and the point where a vertical line through this point would cross a horizontal line connecting the two pupils (mean distance was 86 pixel or 2.7° of visual angle). The end points of the distances were indicated by two white points to make sure that participants knew which distances they had to adjust. The comparison stimulus was presented either horizontally to the right or vertically on bottom of the standard stimulus (Figure 1). The adjusting line was one pixel in width and its initial length was 180 or 20 percent of the distance on the standard stimulus.

The task was explained verbally to the participants. There were 40 trials per experiment: 10 (faces) * 2 (initial line lengths) * 2 (line positions). The

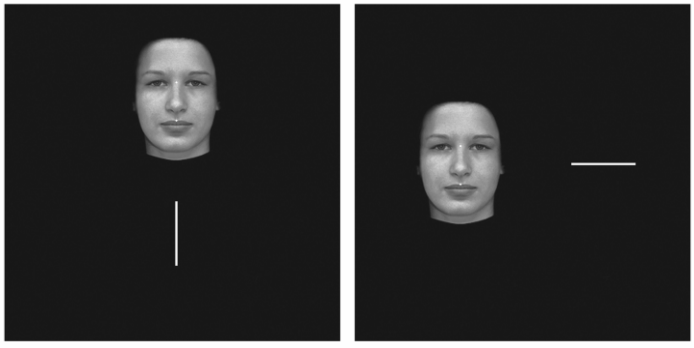

Figure 1. The two possible dispositions of standard and comparison stimulus for one face with the eye-mouth distance indicated with white points as standard stimulus.

order of faces, initial line lengths, and line positions was counterbalanced across participants using a mixed Latin square design.

4.3.2 Results and Discussion

Data were averaged across the two line positions, the two initial line lengths and the ten faces. On average, the eye-mouth distance was overestimated by 34 percent (SEM = 4.57) and the inter-eye distance by 18 percent (SEM = 10.28).

Several researchers have stressed that configural changes in upright faces are very well detectable (Bruce, Doyle, Dench, & Burton, 1991; Haig, 1984; Kemp, McManus, & Pigott, 1990). The large overestimations found in Experiment 1 indicate that this ability does not mean veridical percepts of the configural information.

To investigate several possible explanations for the perceived distances in faces Experiment 2 was conducted with several modifications of one face used in Experiment 1.

4.4 Experiment 2

The aim of Experiment 2 was on the one hand to investigate the role of different facial features (eyes, eyebrows and mouth), and on the other hand to test the role of texture and surface based information onto the perception of distances in faces. Perhaps the oval contour of the head provides a frame of reference that leads to an overestimated percept. This hypothesis was investigated by showing two points in isolation (Figure 2a) vs. the same points within the head context (Figure 2b).

The eyebrows and the mouth might induce Mueller-Lyer like contours (dotted lines in Figure 2d, not shown in the experiment) that could contribute to the large overestimations of the eye-mouth distance. Then, adding the eyes, eyebrows and mouth to the head context (see Figure 2c) would result in larger overestimations compared to the head context alone (Figure 2b).

Furthermore, removing the eyes (Figure 2d) should not affect the perceived eye-mouth distance compared to the whole face (Figure 2f) if the overestimated percept of the eye-mouth distance really is caused by the Mueller-Lyer like induced illusion evoked by the eyebrows and mouth. Note however, that the contours of the eyes might be responsible for a Mueller-Lyer like effect influencing the perceived inter-eye distance (dotted lines in Figure 2c, not shown in the experiment). In this case removing the eyes (Figure 2d) should cause a decrease of the overestimation of the inter-eye distance compared to the whole face.

A further explanation for the overestimation of the eye-mouth distance

could be based on the nose. Perhaps the three dimensional structure of the nose produces an increase of the perceived eye-mouth distance (Figure 2e). A similar prediction would follow from the assumption that the nose introduced an effect comparable to the Oppel-Kundt illusion.

To investigate the role of surface based information onto the perception of distances in faces, the same facial variations were done with a line drawing of the face depicted in Figure 2 (see Figure 3a-f).

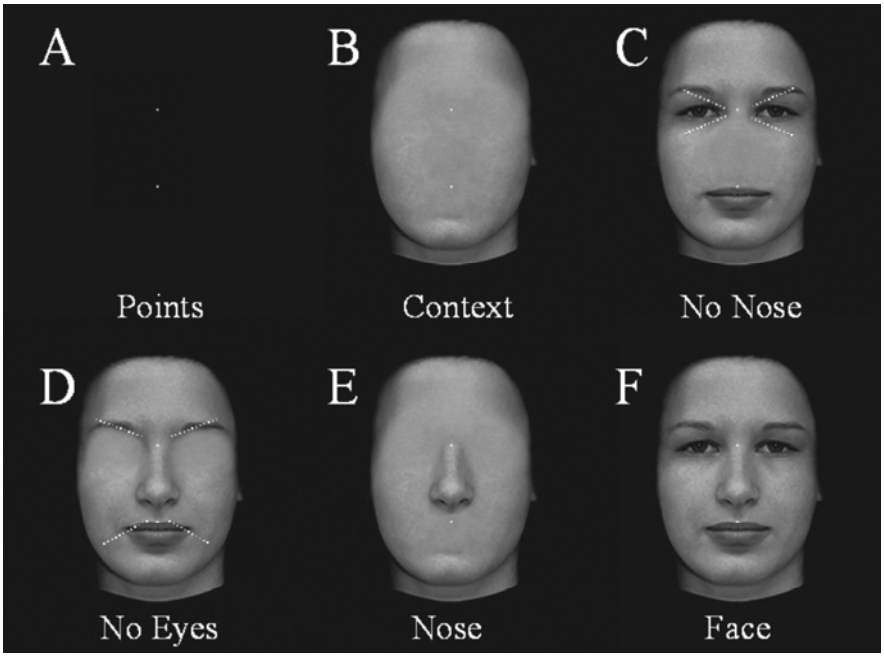

Figure 2. The two white points in the six conditions with the photograph stimuli.

4.4.1 METHOD

4.4.1.1 PARTICIPANTS

Twenty-four undergraduates from the University of Zurich voluntarily participated in this study and were randomly assigned to the two groups of twelve participants. All had normal or corrected to normal vision.

4.4.1.2 MATERIALS AND PROCEDURE

From the 10 faces used in Experiment 1 the face was selected in which the overestimated distances were most similar to the mean overestimation found in Experiment 1. The task was explained verbally to participants and again consisted in adjusting the length of a white line (comparison stimulus) to a standard stimulus. As in Experiment 1, for participant group 1 the standard stimulus was the eye-mouth distance and for participant group 2 this was the inter-eye distance. Two white points indicated the distance, which had to be adjusted with the comparison stimulus. To examine the role of facial features onto the perception of distances in faces six different stimulus types were

created: Two points in isolation (Figure 2a, Points), the two points with the head context (Figure 2b, Context), the two points on the original face without the nose (Figure 2c), the two points on the original face without the eyes but with eyebrows and the mouth (Figure 2d, No Eyes), the points on the head context with the nose as the single feature (Figure 2e, Nose) and the points on the whole unmodified face (Figure 2f, Face). These six different versions were made of the photograph of the face selected from Experiment 1. Then, line drawings of the original photograph and of all different versions were made using Adobe Photoshop 5.5. Figure 3a-f depicts the versions of the line drawings of the same face.

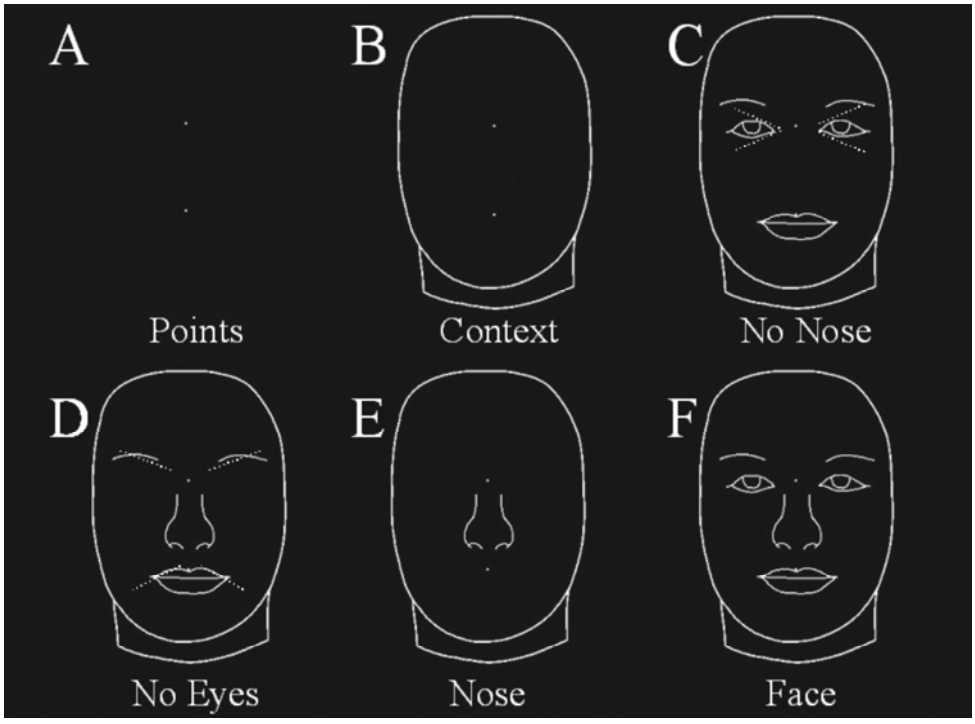

Figure 3. The two white points in the six conditions with the line drawing stimuli.

The adjusting line was one pixel in width and its initial length was 180 or 20 percent of the distance on the standard stimulus. The line was presented either horizontally to the right or vertically beneath of the presented standard stimuli. Participant group 1 adjusted the eye-mouth distance (87 pixel or 2.72° of visual angle) and participant group 2 the inter-eye distance (85 pixel or 2.66° of visual angle). The faces were displayed in three angles of clockwise rotation from upright (0°, 90° and 180°). There were two blocks of 72 trials resulting in 144 trials per experiment: 2 (stimulus type (photograph vs. line drawing)) * 6 (conditions) * 2 (initial length of comparison stimulus) * 2 (positions of standard and comparison stimuli) * 3 (angles of rotation of the standard stimulus).

4.4.2 RESULTS

The data were averaged across the two initial line lengths and the two positions on the screen. The mean overestimation of the eye-mouth distance was 27 percent and the one of the inter-eye distance 16 percent for the upright unmodified photographic face (Figure 2f). For the upright unmodified line drawing face (Figure 3f) the mean constant error of the eye-mouth distance was +21 percent and the one of the inter-eye distance +20 percent. A two-tailed paired samples t-test revealed that for the eye-mouth distance the overestimation was significantly greater in the upright photograph than in the upright line drawing $t(11) = 2.32$, $p < .05$. For the inter-eye distance the overestimation was comparable in both stimulus types $t(11) = -2.02$, $p = .07$.

4.4.2.1 EFFECT OF THE HEAD CONTEXT AND THE EYES, EYEBROWS AND MOUTH (MUELLER-LYER LIKE CONTOURS)

To examine the effects of the head context and the effects of the eyes, eyebrows and the mouth (Mueller-Lyer like contours) onto the perception of distances in faces depending on the stimulus type (photograph, line drawing) we computed a three factor analysis of variance (ANOVA) with stimulus type, condition (with the three different stimuli Points (Figure 2a and 3a), Context (Figure 2b and 3b) and No Nose (Figure 2c and 3c)), and Angle of Rotation (0°, 90°, 180°) as within-participants factors for the eye-mouth distance (participant group 1) and for the inter-eyes distance (participant group 2) separately.

For the eye-mouth distance, there was no main effect of stimulus type $F(1, 11) = 2.15$, $MSE = 38.75$, $p = .17$, but the effects of condition and orientation were significant[1] $F(1.34, 14.74) = 32.47$, $MSE = 439.74$, $p < .001$, and $F(2, 22) = 54.18$, $MSE = 35.42$, $p < .001$, respectively. None of the two twofold interactions with the factor stimulus type reached statistical significance (stimulus type*condition: $F(2, 22) = 1.01$, $MSE = 13.21$, $p = .38$; stimulus type*orientation: $F(2, 22) = 3.36$, $MSE = 17.53$, $p = .05$). The twofold interaction between condition and orientation was significant $F(4, 44) = 6.04$, $MSE = 17.7$, $p < .001$, whereas the threefold interaction between stimulus type, condition and orientation did not lead to a significant value $F(2.49, 27.42) = .97$, $MSE = 36.66$, $p = .41$.

For the inter-eye distance the three factor ANOVA with stimulus type (photograph vs. line drawing), condition (Points, Context, No Nose) and

[1] In all analyses of this study, if Mauchly's (1940) test of sphericity showed a significant deviance (α=0.25) from equicorrelation for a repeated factor or for a combination of factors including at least one repeated factor, Greenhouse and Geisser's (1959) Epsilon was used to adjust the degrees of freedom for the averaged tests of significance.

orientation (0°, 90°, 180°) as within-participants factors also revealed no significant main effect of stimulus type $F(1, 11) = 0.0010$, $MSE = 6.59$, $p = .98$, whereas the effects of condition and orientation again were significant $F(1.39, 15.25) = 25.42$, $MSE = 275.21$, $p < .001$ and $F(1.13, 12.40) = 14.79$, $MSE = 92.75$, $p < .01$, respectively. Neither the interaction between stimulus type and condition $F(2, 22) = 1.56$, $MSE = 21.2$, $p = .23$ nor the interaction between stimulus type and orientation $F(2, 22) = .36$, $MSE = 11.25$, $p = .70$ showed significant values. But the interaction between condition and orientation again reached a significant value $F(4, 44) = 3.22$, $MSE = 10.28$, $p < .05$. There was no significant threefold interaction between stimulus type, condition and orientation $F(2.26, 24.85) = 0.85$, $MSE = 23.89$, $p = .45$.

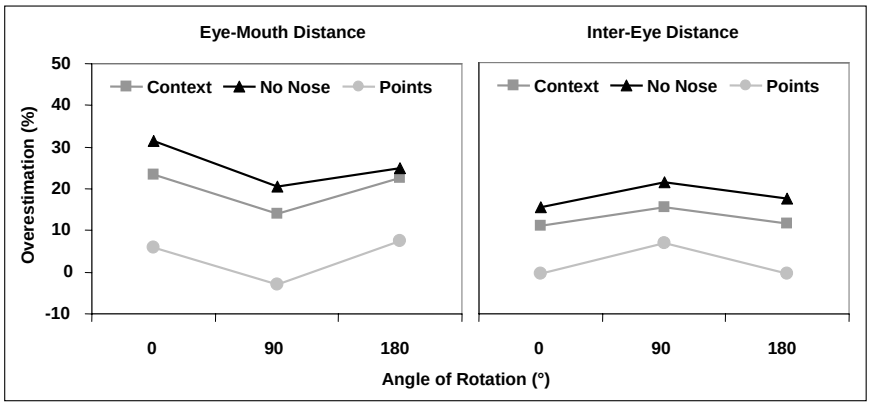

Figure 4. Constant errors in the three conditions Context, No Nose and Points averaged across the two stimulus types photograph and line drawing. Left: Eye-mouth distance. Right: Inter-eye distance.

Because there was neither a main effect of stimulus type nor a significant interaction with stimulus type as one factor for both the eye-mouth distance and the inter-eye distance we averaged the data of the photograph and the line drawing to further analyze the impact of the head context and a possible influence of the Mueller-Lyer like illusion onto the overestimation.

Figure 4 (left) shows the overestimation (constant errors) of the conditions Points, Context and No Nose averaged between the two stimulus types for the eye-mouth distance. It is obvious, that the two points in isolation are perceived almost veridical and that the head context alone already induces an overestimation of the eye-mouth distance of about 20 percent. Furthermore, adding the eyes, eyebrows and the mouth to the head context leads to an increased constant error in all three orientations. Compared to the constant errors in the conditions Context and No Nose the relatively small constant errors in the Points condition reflect approximately the horizontal-vertical illusion (HVI).

A two factor ANOVA with condition (Points, Context, No Nose) and orientation revealed significant main effects of condition $F(1.34, 14.74) = 32.47$, $MSE = 219.87$, $p < .001$, and orientation $F(2, 22) = 54.18$, $MSE = 17.71$, $p < .001$, as well as a significant interaction between condition and orientation $F(4, 44) = 6.04$, $MSE = 8.85$, $p < .01$. Table 1(left) shows the

Bonferroni corrected pairwise comparisons of means for the factor condition.

		Eye-Mouth Distance			Inter-Eye Distance		
(I) Condition	(J) Condition	MD (I-J)	*SE*	*p*	MD (I-J)	*SE*	*p*
Context	No Nose	-5.45	1.58	.02	-5.22	1.36	.01
Context	Points	16.67	3.21	.00	10.87	2.52	.00
No Nose	Points	22.12	3.43	.00	16.09	2.77	.00

Table 1. Bonferroni corrected pairwise comparisons between the three conditions Points, Context and No Eyes used in Experiment 2 (based on estimated marginal means).

As can be seen the constant error was significantly greater when the head context was added to the two points. If the eyes, eyebrows and mouth were additionally added to the head context the overestimation of the eye-mouth distance got even larger. The Bonferroni corrected pairwise comparisons for orientation showed that the constant errors were significantly smaller at 90° than at 0° ($p < .001$) and 180° ($p < .001$).

If the eyebrows and the mouth really induce a Mueller-Lyer like illusion for the perceived eye-mouth distance, a face without eyes would evoke comparable overestimations of the eye-mouth distance as an unmodified face. Computing a three factor ANOVA with stimulus type (photograph vs. line drawing), condition (Face, No Eyes) and orientation (0°, 90°, 180°) as within participants-factors for the eye-mouth distance revealed a significant main effect of stimulus type $F(1, 11) = 8.62$, $MSE = 45.51$, $p < .05$, no significant effect of condition $F(1, 11) = 0.04$, $MSE = 16.34$, $p = .84$, whereas the effect of orientation was significant once again $F(1.35, 17.87) = 11.32$, $MSE = 63.22$, $p < .001$. No significant interaction could be revealed (stimulus type*condition: $F(1, 11) = 0.93$, $MSE = 24.14$, $p = .36$; stimulus type*orientation: $F(2, 22) = 1.05$, $MSE = 37.88$, $p = .37$; condition*orientation: $F(2, 22) = 2.93$, $MSE = 9.73$, $p = 0.7$; stimulus type*condition*orientation: $F(2, 22) = 1.71$, $MSE = 20.87$, $p = .20$). The two separate two factor ANOVAs for the photograph and line drawing with condition (Face, No Eyes) and orientation (0°, 90°, 180°) as within-participants factor for the eye-mouth distance showed no significant differences between the unmodified face and the face without the eyes neither for the photograph $F(1, 11) = 0.83$, $MSE = 18.59$, $p = .38$ nor for the line drawing $F(1, 11) = 0.35$, $MSE = 21.89$, $p = .57$. Thus, the overestimations of the eye-mouth distance are comparable in the unmodified face and the face without the eyes.

Figure 4 (right) shows the constant errors for the inter-eye distance

averaged for the photograph and the line drawing. As for the eye-mouth distance adding the head context alone already leads to an overestimation. Furthermore, adding the eyes leads to an enlargement of the overestimation.

A two factor ANOVA with condition (Points, Context, No Nose) and orientation as within participants factor taken together for the two different stimulus types revealed significant main effects of condition and orientation, $F(1.39, 15.25) = 25.42$, $MSE = 137.61$, $p < .001$ and $F(1.13, 12.40) = 14.79$, $MSE = 46.38$, $p < .01$, respectively. There was also a significant interaction between condition and orientation $F(3.40, 37.43) = 3.22$, $MSE = 6.04$, $p < .05$ (Figure 4, right).

The Bonferroni corrected comparisons for condition again showed significant differences between the conditions Context and Points, Context and No Nose and between Points and No Nose (Table 1, right). The Bonferroni corrected comparisons for orientation revealed significant differences between 0° and 90° ($p < .01$) and between 90° and 180° ($p < .05$).

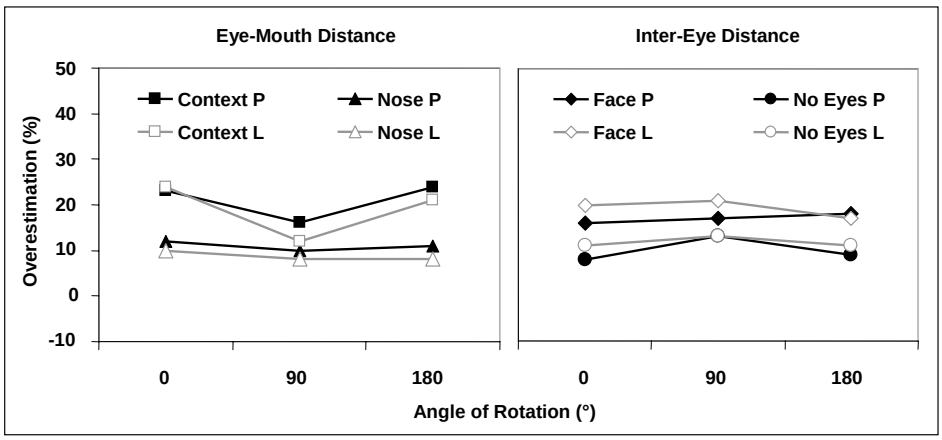

Figure 5. Constant errors in the two conditions Context and Nose for the eye-mouth distance (left) and the two conditions Face and No Eyes for the inter-eye distance (right), separately for the photographs and for the line drawings.

If the contours of the eyes really induce a Mueller-Lyer like illusion for the inter-eye distance, a face without any eyes (Figure 2d and 3d) should lead to a smaller overestimation than an unmodified face. Figure 5 (right) shows the constant errors for the unmodified face and the face without any eyes for the two stimulus types separately. As can be seen, the overestimations are smaller in the face without any eyes compared to the unmodified face.

A three factor ANOVA with stimulus type (photograph, line drawing), condition (Face, No Eyes) and orientation (0°, 90°, 180°) revealed a main effect of stimulus type $F(1, 11) = 6.02$, $MSE = 31.74$, $p < .05$ and condition $F(1, 11) = 19.60$, $MSE = 86.10$, $p < .01$ but no effect of orientation $F(2, 22) = 1.90$, $MSE = 47.27$, $p = .17$. Neither the interaction between stimulus type and condition $F(1, 11) = 0.56$, $MSE = 21.21$, $p = .47$, nor that between stimulus type and orientation $F(1.54, 16.91) = 1.49$, $MSE = 23.08$, $p = .25$, nor the interaction between condition and orientation $F(2, 22) = 0.84$, $MSE = 19.41$, $p = .45$ showed significant values. The threefold interaction between

stimulus type, condition and orientation was not significant either $F(1.37, 15.12) = 1.33$, $MSE = 45.26$, $p = .28$. Both of the two separate two factor ANOVAs for each stimulus type (photograph, line drawing) with condition (Face, No Eyes) and orientation (0°, 90°, 180°) as within factor revealed a significant effect of condition for the photograph $F(1, 11) = 9.19$, $MSE = 77.05$, $p < .05$ as well as for the line drawing stimulus $F(1, 11) = 32.74$, $MSE = 30.26$, $p < .001$.

4.4.2.2 EFFECT OF THE NOSE ON THE EYE-MOUTH DISTANCE

Figure 5 (left) shows the effect of the nose onto the perceived eye-mouth distance for the photograph and the line-drawing separately in the three orientations. Obviously, the overestimation is smaller in upright and inverted faces when the nose is added to the head context. This challenges the hypothesis that the nose induces an Oppel-Kundt like illusion.

A three factor ANOVA with stimulus type (photograph vs. line drawing), condition (with the two stimuli Context (Figure 2b and 3b) and Nose (Figure 2e and 3e)) and orientation (0°, 90°, 180°) as within-participants factors showed significant main effects of stimulus type $F(1, 11) = 7.43$, $MSE = 23.92$, $p < .05$, condition $F(1, 11) = 108.84$, $MSE = 34.55$, $p < .001$ and orientation $F(2, 22) = 12.50$, $MSE = 34.23$, $p < .001$. There was no significant interaction with the factor stimulus type (stimulus type*condition: $F(1, 11) = 1.09$, $MSE = 11.5$, $p = .76$; stimulus type*orientation $F(2, 22) = 1.29$, $MSE = 16.24$, $p = .30$; stimulus type*condition*orientation $F(2, 22) = 1.46$, $MSE = 12.33$, $p = .25$). The twofold interaction between condition and orientation was significant $F(2, 22) = 9.35$, $MSE = 23.88$, $p < .01$ (Figure 5, left). Bonferroni corrected pairwise comparisons revealed that the mean constant error was significantly smaller at 90° than at 0° ($p < .01$) and 180° ($p < .05$).

4.4.3 DISCUSSION

The goal of Experiment 2 was to investigate the role of facial features onto perceived distances in faces. A first explanation based on a geometrical context effect was consistent with the results. When the distance between two points had to be estimated, adding an oval head context (without the eyes, eyebrows, mouth, and nose) lead to an overestimation of 23.5 percent[2] (eye-mouth distance) and 11.0 percent (inter-eye distance) in upright orientation. Adding the eyes, eyebrows and mouth to the context increased the overestimation of the eye-mouth distance by 8 percent and the inter-eye

[2] If not explicitly mentioned, the indicated overestimations are averaged across the two upright stimulus types, photograph and line drawing.

distance by 4.5 percent. Some contours of the eyes, eyebrows and the mouth might have introduced an effect similar to the Mueller-Lyer illusion (see Figure 2c and 3c). Removing the eyes did not affect the overestimation of the eye-mouth distance compared to the whole face, but reduced the perceived inter-eye distance by about 8 percent compared to the whole face, which is compatible with an explanation based on Mueller-Lyer like effects.

Two other hypotheses for the large overestimations found in Experiment 1 were tested in the condition in which the nose was added to the oval head context (without eyes, eyebrows, and mouth). Compared to the head context alone this manipulation decreased the perceived eye-mouth distance between 11 (photograph) to 14 percent (line drawing). Therefore, neither an explanation based on the Oppel-Kundt illusion, nor the hypothesis that the three dimensional information of the nose would lead to an increased percept of the eye-mouth distance could be confirmed. This conclusion is also supported by the fact that the eye-mouth distance was overestimated about 15 percent (photograph) and about 11 percent (line drawing) more in the intact face than in the condition in which only the nose was added to the head context. Similarly, the eye-mouth distance was less overestimated (15 percent) in the intact face than in the face without a nose.

A further aim of Experiment 2 was to investigate the role of surface based information onto the perceived configural information in faces. Using versions of line drawings we found similar results for the line drawings as for the photographic stimuli. Thus, a main influence of texture and surface based information on perceived distances in faces could be ruled out. Instead, our results suggest that the large overestimations are related to classical geometrical illusions.

4.5 General Discussion

In this study the perception of configural information in faces and possible influences of facial features onto these percepts were examined. Several studies which have investigated the ability of detecting configural manipulation in faces have revealed that human observers are very sensitive to such changes (Bruce, Doyle, Dench, & Burton, 1991; Haig, 1984; Kemp, McManus, & Pigott, 1990). This could lead to the assumption that we not only are very sensitive to configural changes in upright faces but also that we have very precise percepts of configural information in faces. Interestingly this is not the case at all. In contrast, we found in a previous study (Schwaninger et al., 2003) and in Experiment 1 of this study large overestimations of the eye-mouth distance (previous study: 41 percent, Experiment 1: 34 percent) and inter-eye distance (previous study: 11 percent,

Experiment 1: 18 percent). These overestimations are larger than any known perceptual size illusions (Coren & Girgus, 1978).

Additionally, we were interested in the role of facial features and texture for these large overestimations. Several possible influences of different features were tested. First, the results suggest that there is an influence of the frame of reference provided by the head context. Second, the results also show that the eye-mouth distance, as well as the inter-eye distance are perceived greater when features which induce a Mueller-Lyer like illusion (eyebrows and mouth for the eye-mouth distance, eyes for the inter-eye distance) are added to the head context. Third, a possible influence of an Oppel-Kundt like illusion or of depth information which would result in an overestimation of the eye-mouth distance can be rejected because adding the nose onto the head context *decreased* the perceived eye-mouth distance.

In sum, these findings suggest an important role of well-known perceptual illusions for the explanation of the large overestimation of configural information in faces.

4.6 REFERENCES

Bruce, V., Doyle, T., Dench, N., & Burton, M. (1991). Remembering facial configurations. *Cognition, 38*, 109-144.

Coren, S., & Girgus, J.S. (1978). *Seeing is Deceiving: The Psychology of Visual Illusions*. Hillsdale, NJ: Erlbaum.

Fick, A. (1851). *De errore quodam optico asymenetria bulbi effecto*. Marburg: J.A. Kochii.

Haig, N.D. (1984). The effect of feature displacement on face recognition. *Perception, 13*, 505-512.

Kemp, R., McManus, C., & Pigott, T. (1990). Sensitivity to the displacement of facial features in negative and inverted images. *Perception, 19*, 531-543.

Künnapas, T. M. (1955). Influence of frame size on apparent length of a line. *Journal of Experimental Psychology, 30*, 168-170.

Künnapas, T.M. (1957a). The vertical-horizontal illusion and the visual field. *Journal of Experimental Psychology, 53*, 405-407.

Künnapas, T.M. (1957b). Vertical-horizontal illusion and surrounding fields. *Acta Psychologica, 13*, 35-42.

Kundt, A. (1863). Untersuchungen ueber Augenmass und optische Taeuschungen. *Poggendorff Analle, 120*, 118-158.

Leder, H., & Bruce, V. (2000). When inverted faces are recognized: the role of configural information in face recognition. *Quarterly Journal of Experimental Psychology Section A-Human Experimental Psychology, 53*, 513-536.

Mauchly, J.W. (1940). Significance test for sphericity of a normal n-variate distribution. *Annals of Mathematical Statistics, 11*, 204-210.

Mueller-Lyer, F.C. (1889). Optisches Urteilstäuschungen. *Dubois-Reymonds Archive für Anatomie und Physiologie (Suppl.)*, 263-270.

Murray, J.E., Yong, E., & Rhodes, G. (2000). Revisiting the perception of upside-down faces. *Psychological Science, 11*, 492-496.

Oppel, J.J. (1855). Ueber geometrisch-optische Taeuschungen. *Jahresbericht des Frankfurter Vereins*, 37-47.

Schwaninger, A., & Mast, F. (1999). Why is face recognition so orientation-sensitive? Psychophysical evidence for an integrative model. *Perception, 28 (Suppl.)*, 116.

Schwaninger, A., Ryf, S., & Hofer, F. (2003). Configural information is processed differently in perception and recognition of faces. *Vision Research, 43*, 1501-1505.

Searcy, J.H., & Bartlett, J.C. (1996). Inversion and processing of components and spatial relational information in faces. *Journal of Experimental Psychology: Human Perception and Performance, 22*, 904-915.
Sergent, J. (1984). An investigation into component and configurational processes underlying face recognition. *British Journal of Psychology, 75*, 221-242.
Yin, R.K. (1969). Looking at upside-down faces. *Journal of Experimental Psychology, 81*, 141-145.

5 THE ROLE OF CO-OCCURRENCE FOR VIEW-BASED OBJECT RECOGNITION

5.1 ABSTRACT

Many object recognition theories suppose a strictly serial and bottom-up processing and disregard top-down feedback from later stages to earlier ones (for an overview see for example Schwaninger, 2004; Tarr & Bülthoff, 1998). In contrast, current research suggests recurrent feedback during the processing of an object (e.g., Bar, 2003; Graboi & Lisman, 2003; Humphreys, Riddoch, & Price, 1997). Further evidence for top-down influences in object recognition using a semantic priming approach is presented in this study. In Experiment 1 participants had to name objects in two different viewpoints (canonical vs. non-canonical), which were either preceded by a contextual consistent or a contextual inconsistent priming stimulus. We found clear effects of prime consistency and target viewpoint, and a significant interaction between prime consistency and target viewpoint. In Experiment 2 we could replicate these findings with a non-linguistic contextual association task. Presenting a consistent prime prior to a target reduced the viewpoint dependency significantly. Different models of top-down influences, which could explain these effects, are discussed within a view-based object recognition framework.

5.2 INTRODUCTION

Many object recognition theories are bottom-up and serial in nature. The two main approaches of object recognition theories are well known examples. On the one hand object-centered models, e.g., the traditional approach of Marr (1982) or the recognition-by-components theory by Biederman (1987) and Hummel and Biederman (1992) assume object-centered representations and propose that objects are stored as descriptions of spatial arrangements among parts within an object-centered 3D-coordinate system. According to this approach such an object-centered representation specifies the parts and their inter-relationships and results in viewpoint-independent object recognition. In contrast to this object-centered approach the view-based approaches assume view-specific representations of objects (e.g., Bülthoff & Edelman, 1992; Tarr & Bülthoff, 1995; Riesenhuber & Poggio, 1999; for a review see Edelman, 1999; Tarr & Bülthoff, 1998; Ullman, 1996) and therefore postulate viewpoint-dependent object recognition.

There are many empirical findings which confirm the viewpoint-dependent object recognition. For example orientation effects were found for novel objects (e.g., Bülthoff & Edelman, 1992; Edelman & Bülthoff, 1992;

Tarr & Pinker, 1989), and also for common, familiar objects (e.g., Hayward & Tarr, 1997; Lawson & Humphreys, 1996, 1998; Murray, 1997, 1999; Newell & Findlay, 1997; Palmer, Rosch & Chase, 1981). Orientation-dependent performance could also be verified with naming tasks, sequential matching and priming tasks (for review see Jolicoeur & Humphrey, 1998), with a visual search task (Jolicoeur, 1992) and also with figure-ground segregation tasks (Gibson & Peterson, 1994). Orientation-dependent recognition performance is not limited to individual objects, like faces (e.g., Hill, Schyns & Akamatsu, 1997), or to objects on the subordinate level of categorization (e.g., Edelman & Bülthoff, 1992; Tarr, 1995), but is also true for basic level recognition (Hayward & Williams, 2000; Lawson & Humphreys, 1998; Murray, 1998; Palmer, Rosch & Chase, 1981). Recently Graf, Kaping, and Bülthoff (2005) proposed that object recognition is based on coordinate transformations which compensate for spatial transformations. According to that approach, not the internal stimulus is aligned, but the perceptual coordinate system (reference frame) is adjusted, so that correspondence is achieved between positions specified in memory and positions in the visual fields. If this is true, this internal coordinate system could be active for some time and would then lead to a facilitated object recognition of subsequently presented objects in the same orientation (orientation congruency effect), even when their shapes differ remarkably. This is exactly what the authors found.

Furthermore, recent studies started to investigate how certain characteristics of structural description models and view-based models could be combined in the object recognition process (see also Hayward, 2003 on this topic). For example, Foster and Gilson (2000) propose a simple model with two independent terms, one represents object structure (independent of viewpoint) and the second image-based features (independent of structure). With this additive model they can explain why object recognition is viewpoint dependent, but simultaneously sensitive to non-accidental features.

The classical object recognition theories disregard influences of top-down processes, like for example expectancies and knowledge. But in real world our expectation and our knowledge has an omnipresent influence onto our perception, which cannot be negated. For example, if one enters one's own office one would be extremely astonished if the computer monitor wasn't at the place where it used to be or if its color had changed suddenly.

Before summarizing the most relevant experimental findings in literature about such top-down influences it is worth noting that top-down processing can be defined in two different ways. Psychophysically one can define top-down influence as 'influence of previously activated representation onto the processing of subsequent stimuli'. In a more general way this means nothing

else than that our knowledge and expectancies influence what we perceive (e.g., Kosslyn, 1994). A strictly neurophysiologic definition terms top-down processing as recurrent feedback from higher (visual) brain areas to lower (visual) brain areas (e.g., Kosslyn, 1994; Mumford, 1992; Ullman, 1995, 1996). In the model of Ullman (1995) recognition of an object involves bottom-up as well as top-down streams, so that in a bidirectional process many expectations can be explored simultaneously. Bar (2003) proposes a detailed mechanism for the activation of such top-down facilitation during object recognition: First, low-spatial frequency information is projected from early visual areas to the prefrontal cortex, where several expectations about possible interpretations of the input are activated. These "initial guesses" are then back-projected to the infero-temporal cortex and are integrated in the bottom-up process.

In addition, Graboi & Lisman (2003) propose a further neurophysiologic model, which is also psychophysically plausible. They show in their hierarchical model how an attention-based recognition process could be organized by the parallel processing of top-down and bottom-up streams. This parallel processing could move attention to information rich regions and therefore could lead to a very efficient recognition process.

Empirical evidence for top-down influences onto the recognition of objects comes for example from a study of Liter and Bülthoff (1997). They report viewpoint-independency in a name verification task compared to a simple naming task. The authors found increased naming latencies for non-canonical views compared to canonical views, which is consistent with the central assumptions of the viewpoint-dependent models of object recognition (e.g., Bülthoff & Edelman, 1992; Tarr & Bülthoff, 1995; Riesenhuber & Poggio, 1999; for a review see Edelman, 1999; Tarr & Bülthoff, 1998; Ullman, 1996). Interestingly no effect of viewpoint was found when the participants had to verify the name of an object, i.e. if a name matches the visually presented object. An explanation of the authors is that the presentation of a name could activate a multiple views representation and therefore could "pre-activate" different stored views of this object. Such a 'pre-activation' could be especially helpful for recognizing objects in non-canonical views, because it can help to minimize the confusion with another object that has similar viewpoint-specific descriptions (Liter & Bülthoff, 1997). Similar to this, Takano (1989) suggests that strategic processes or expectancies about the stimulus can reduce the viewpoint-dependence during object recognition. In addition, Wilson and Farah (2003) found empirical evidence that top-down strategic processes can influence the ability to recognize orientation-invariant local features of objects.

Further evidence for top-down influences comes for example from studies of Biederman et al., (1982), Biederman et al., (1988), Hollingworth & Henderson (1998) and Palmer (1975). The main finding of these studies is that the identification of an object is affected by the context in which it is presented. The context can apparently facilitate the identification of an object but also can have the opposite effect: Inconsistent context can even hamper the identification of an object (Biederman et al., 1982). Not only the context influences the recognition/identification of objects but also associated objects can facilitate subsequent object identification (e.g., Henderson, Pollatsek, & Rayner, 1987). Similar to Liter and Bülthoff (1997) these authors assume that objects prime the internal representations of associated objects. Boyce, Pollatsek, and Rayner (1989) emphasize that the episodic relatedness of objects (the information whether or not it is plausible that two objects co-occur in the same context) could be stored in associative memory and could drive this kind of top-down priming.

In the present study we further investigated the influence of co-occurrence of episodic related objects onto the identification of subsequent stimuli using a naming task (Experiment 1) and a contextual association task (Experiment 2). The main goal of this study was to investigate the influence of co-occurrence of episodic related objects onto the viewpoint-dependency of object identification. Instead of showing the name prior to the object that has to be recognized (name verification) a more natural condition would be to show a preceding object, which is associated with the target object. For example tea spoons tend to be near to tea cups. Thus, it could be expected that looking at a tea spoon activates the viewpoint-specific descriptions of a tea cup, which would help to recognize it even in unusual views.

5.3 Experiment 1

5.3.1 Method

5.3.1.1 Participants

Twelve undergraduate students from the University of Zurich (6 females, 6 males) voluntarily took part in Experiment 1. All participants reported normal or corrected-to-normal vision. They were all naive with regard to the hypotheses under investigation.

5.3.1.2 Materials And Procedure

A selection of 64 colored, common objects was chosen as stimuli. The objects were selected from a list of 3D Studio Max objects. Half of the objects served as priming stimuli, whereas the other half was defined as target stimuli.

Of each target stimulus, a canonical and a non-canonical view were used (see Figure 1 and Appendix for canonical and non-canonical views used in Experiment 1). For the priming stimuli, only one view was used and it was chosen to be about 45 degrees rotated away from the view of the target stimuli. For details on the concept of a canonical view see for example Blanz, Tarr, & Bülthoff (1999).

The experiments were conducted in a dimly lit room. The viewing distance was maintained by a head rest so that the center of the screen was at eye height of the participants and the mean average visual angle of the stimuli

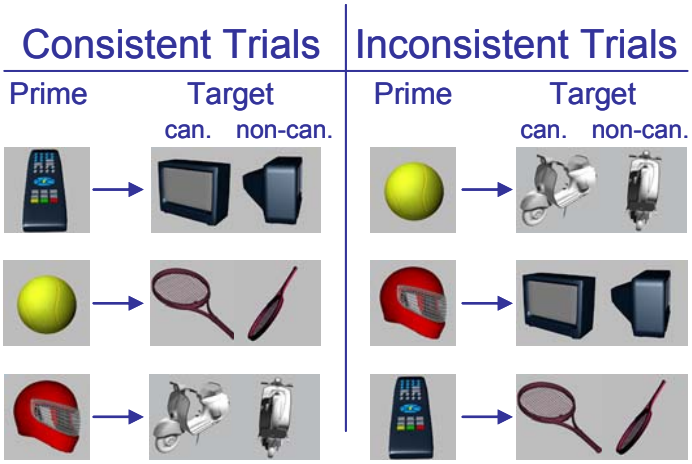

Figure 1. Examples of stimuli used in this study. Left: Consistent Trials = prime and target objects are episodically related (can. = canonical view, non-can. = non-canonical view), Right: Inconsistent Trials = prime and target are not related episodically.

covered 6° (distance of the participant to the screen was 60 cm). All objects were presented with the experimental software E-Prime Version 1.0 on a 15-inch screen on a gray background. In a naming task participants had to name each presented object as quickly and accurate as possible. The time between the onset of the presentation of an object and the voice onset time was recorded (=naming latency) by a Voice Key (Serial Response Box for E-Prime (SRBox)). One trial consisted of a fixation cross for 1000 ms followed by the first object (priming stimulus) to be named. After this object had been named, a masking stimulus was shown for 1000 ms and then the target object had to be named. In half of the trials the two sequentially presented objects tended to co-occur in real world, so that they were associated - e.g., the first object (priming stimulus) was a remote control and the second object (target stimulus) depicted a television. In the other half of the trials the objects were not contextually related (e.g., remote control and dog).

Additionally each target stimulus was once presented in the canonical view and once in a non-canonical view. There were 4 blocks of 32 trials each. In each block all 32 target stimuli were presented randomly and the experimental conditions (consistent/canonical, inconsistent/non-canonical, consistent/non-canonical, inconsistent/canonical) were counterbalanced so that in one block each condition occurred 8 times.

There was a total of 128 trials per Experiment: 2(consistent vs.

inconsistent priming stimulus) * 2(canonical vs. non-canonical target view) * 32(target objects).

Prior to the Experiment there were 8 practice trials in which each experimental condition occurred 2 times.

5.3.2 RESULTS

7.4 percent (114 out of a total of 1536 responses) of the data were naming errors. These naming errors were excluded from the statistical analysis. Additionally, to eliminate outliers we disregarded naming latencies greater than the mean naming latency (M = 967.16 ms) plus 2 standard deviations (SD = 627.09 ms). Thus 86 percent (1321 trials of 1536 responses) of all responses were analyzed.

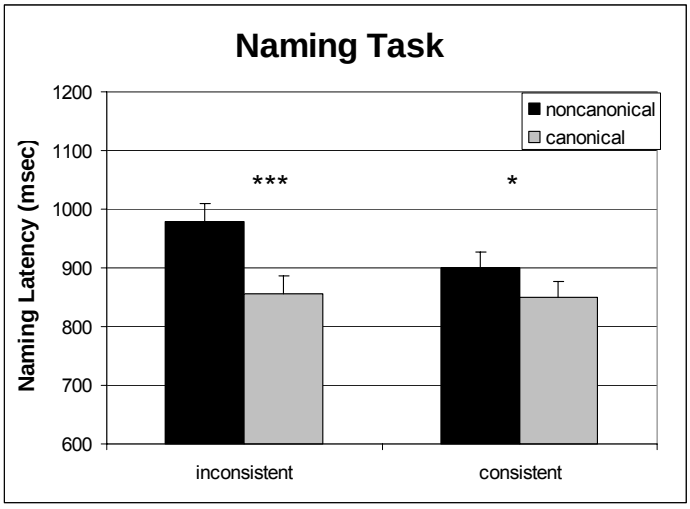

Figure 2. Naming latencies (ms) in the four conditions of Experiment 1. * $p < .05$, *** $p < .001$.

Naming latencies in the four experimental conditions can be seen in Figure 2. Presenting a consistent prime reduced the viewpoint-dependency of the naming of the target. A two factor analysis of variance (ANOVA) with consistency (consistent vs. inconsistent prime) and view (canonical vs. non-canonical target-view) as within-participants factor was carried out. There were reliable main effects of consistency, $F(1, 11) = 2310.66$, $MSE = 0.462$, $p < .05$, and view $F(1,11) = 1579.26$, $MSE = 0.838$, $p < .001$, as well as a significant interaction between consistency and view $F(1, 11) = 1550.99$, $MSE = 0.491$, $p < .01$.

The separate one factor analysis of variance (ANOVA) with view as within-participants factor revealed significant differences between the two views for both the consistent and inconsistent trials, $F(1, 11) = 7.16$, $MSE = 2042.97$, $p < .05$; $F(1, 11) = 84.04$, $MSE = 1087.28$, $p < .001$, respectively. To analyze the effect of prime consistency onto canonical and non-canonical views one factor analysis of variance (ANOVA) with consistency as within-participants factor were carried out for the two target views. The effect of consistency was significant only for the non-canonical views $F(1, 11) = 13.87$, $MSE = 2744.54$, $p < .01$.

5.3.3 DISCUSSION

The viewpoint-dependency of the target objects was reduced significantly if the target object was preceded by an episodic associated object. The reliable

interaction between consistency and view type indicates that the co-occurrence of related objects can reduce the viewpoint-effect. One possible explanation for the facilitated object identification could be that a 'pre-activation' of episodic related objects in different views could be especially helpful for recognizing them in non-canonical views, because it can help to resolve the competition with another object that has similar viewpoint-specific descriptions (Liter & Bülthoff, 1997).

The results of Experiment 1 support the hypothesis that object identification is not strictly driven bottom-up but can be facilitated by top-down processes induced by co-occurrence of objects. Note that several other researches have pointed out the role of such top-down processes in object recognition and identification (for a review see Henderson & Hollingworth, 1999 or Kosslyn, 1994). The main new finding of Experiment 1 however is that an associated object reduces the viewpoint-dependency of the naming of another object, which could be explained by a multiple view representation, which can be 'pre-activated' by a semantically associated object and thus facilitate the recognition process.

In Experiment 2 we investigated whether the same effect can be found in a non-linguistic task. Therefore, we used a contextual association task in which the participants had to decide as accurately and fast as possible if two sequentially presented objects tend to co-occur in the real world.

5.4 EXPERIMENT 2

5.4.1 METHOD

5.4.1.1 PARTICIPANTS

12 undergraduates from the University of Zurich (6 females, 6 males) participated in Experiment 2. All participants reported normal or corrected-to-normal vision. They were naive with regard to the hypotheses under investigation.

5.4.1.2 MATERIALS AND PROCEDURE

The apparatus was identical to Experiment 1. Because in Experiment 1 the total error rate was larger than 10 percent we replaced all objects of Experiment 1 for which the naming error rate was greater than 10 percent by other objects. Additionally viewpoints of the target stimuli were changed if the naming error rate was larger than 10 percent. Thus, the stimuli were the same as in Experiment 1 with the exception of the replaced objects and changed viewpoints. Two prime and two target objects were replaced. Additionally the viewpoint was changed for three target objects.

A contextual association task was used. Participants had to decide as

quickly and accurately as possible if the second (target stimulus) of two sequentially presented objects tends to co-occur with the first one (priming stimulus) in real world by pressing one of two buttons on a mouse device. Stimulus-to-response button assignments were counterbalanced across block order and participants. Following this, participants rated on a scale from 0 (prime and target never co-occur in the same context) to 9 (prime and target always co-occur in the same context) how often the prime and target objects co-occur in the same context.

The same experimental conditions were used as in Experiment 1. Half of the trials (64 trials) were consistent trials (prime and target object tend to co-occur) whereas the remaining trials contained non-associated objects. The target objects again were shown in both canonical and non-canonical views.

The principal experimental design was the same as in Experiment 1. Again, there were 4 blocks of 32 trials. In each block all 32 target stimuli were presented randomly and the four experimental conditions (consistent/canonical, inconsistent/non-canonical, consistent/non-canonical, inconsistent/canonical) were counterbalanced so that in one block each condition occurred 8 times. Thus, there were 128 trials per experiment: 2(consistent vs. inconsistent priming stimulus) * 2(canonical vs. non-canonical view) * 32 (target objects).

Again prior to the Experiment there were 8 practice trials in which each experimental condition occurred 2 times.

5.4.2 RESULTS

4.95 percent (76 out of a total of 1536 responses) of the data were errors. Additionally, because the distribution of reaction times was biased to lower reaction times we disregarded naming latencies longer than the mean naming latency (M = 896.41 ms) plus 2 standard deviations (SD = 420.38). Thus 91.47 percent (1405 trials of 1536 responses) of all responses were analyzed.

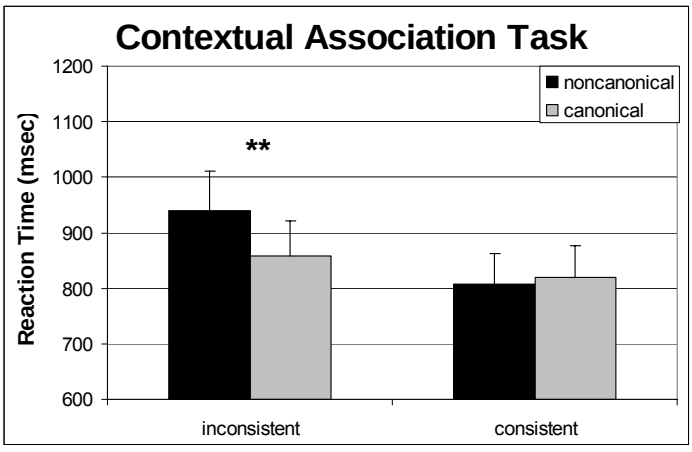

Figure 3. Reaction times (ms) in the four conditions for experiment 2. ** $p < .01$.

The reaction times in the four experimental conditions can be found in Figure 3.

A two factor analysis of variance (ANOVA) with consistency (consistent vs. inconsistent) and view as within-participants factors was carried out. There were reliable main effects of consistency, $F(1, 11) = 6.46$, $MSE = 13751.79$, $p < .05$,

view $F(1, 11) = 7.43$, $MSE = 1927.17$, $p < .05$, and there was a significant interaction between consistency and view $F(1, 11) = 19.30$, $MSE = 1382.62$, $p < .01$.

The separate one factor analysis of variance (ANOVA) with view as within-participants factor for the consistent and inconsistent trials revealed significant differences only for the inconsistent trials $F(1, 11) = 17.12$, $MSE = 2339.78$, $p < .01$. To analyze the effect of prime consistency onto canonical and non-canonical views one factor analysis of variance (ANOVA) with consistency as within-participants factor were carried out for the two target views. There was a significant effect of consistency for non-canonical views, $F(1, 11) = 13.02$, $MSE = 8178.08$ $p < .05$.

Figure 4 shows a histogram of the ratings made by the participants, how often the priming and target stimuli co-occur in the same context (0 = never, 9 = always). As can be seen, most of the consistent trials were also rated as being consistent (5-9 on the rating scale), whereas the co-occurrence of the pairs of stimuli of the inconsistent trials was judged to be rather low (0-4 on the rating scale). This result supports the validity of the used concept "consistence" between the priming and target stimuli.

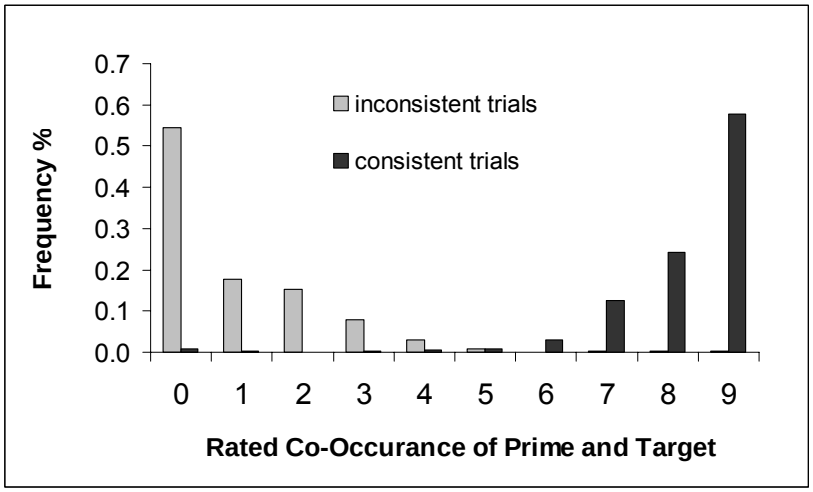

Figure 4. Histogram of the co-occurrence ratings of Experiment 2; 0 means "target and prime never co-occur in the same context", 9 stands for "target and prime always occur in the same context".

5.4.3 Discussion

In Experiment 2 we could replicate the results of Experiment 1 using a contextual association task. Presenting an episodic related object facilitates the recognition of a subsequent object in terms of a reduction of the viewpoint-dependency. Again, the interaction between prime consistency and target view indicates that top-down driven expectancies can facilitate the recognition process of objects in non-canonical views.

5.5 General Discussion

It is well known that the recognition of an object is not only driven bottom-up but can be influenced through top-down processes (e.g., Bar,

2003; Biederman et al., 1982; Biederman et al., 1988; Hollingworth & Henderson, 1998; Liter & Bülthoff 1997; Palmer, 1975, Wilson & Farah, 2003). The main new result of this study is the empirical finding that this top-down influence, not only facilitates the recognition of an object in general (known as general priming effect), but is especially helpful for the recognition of objects in non-canonical views. A preceding object can reduce the viewpoint-dependent recognition of a following contextual associated object. This was shown in a linguistic (naming task) as well as in a non-linguistic task (contextual association task). The finding that top-down influence is especially helpful for the recognition of objects in non-canonical views is consistent with the conclusion of Bar (2003), that the top-down facilitation is more pronounced when recognition is difficult (e.g., brief and masked presentations, low contrast or occluded objects) than when recognition is very easy.

Before accepting the explanation of a top-down processing, however, the two different definitions of top-down explained in the introduction must be addressed. If the concept top-down is just defined as 'influence of knowledge and expectancies' clear top-down influences could be found in this study. However, using the neurophysiologic definition, the finding of this study then is not easily attributable to top-down influences, as will be illustrated in the following paragraph.

There is ample evidence for a hierarchical processing in object recognition. For example Riesenhuber & Poggio (1999) and Knoblich, Riesenhuber, Freedman, Miller, & Poggio, (2002) developed a hierarchical neuronal model of object categorization, which is built up of different layers with different properties. Convergent connection cells in associative memory function as categorization units, which are able to differentiate between objects. Another hierarchical model is known from Perrett, Oram, & Ashbridge (1998) for face recognition, which can explain the viewpoint dependent recognition of faces by neuronal fine and broadly tuned face-view units. Thinking in such a hierarchical model, the finding that an object can reduce the viewpoint-dependency of an associated object could be explained in different ways. One possibility is that the priming occurs through lateral connections within associative or visual memory. For example, the preceding priming stimulus could activate its representation in associative memory, for example its name. This activation could facilitate the activation of the name of the episodic associated object through lateral connections. This "lateral" priming mechanism could be similar to the semantic priming mechanism, which leads to a facilitated word recognition if participants know the category of the words in advance (Lorch, Balota, & Stamm, 1986; Neely, 1991; Neely, Keefe,

& Ross, 1989). The recognition of the associated non-canonical object thus would be fastened by its facilitated access of the name in associative memory. The second possible mechanism is that the visual representation of the priming stimulus itself could directly 'pre-activate' the multiple-view representations of the associated objects through lateral connections in the visual memory. Note however that a model that uses only coordinate transformations to cope with different viewpoints (Graf et al. 2005) would fail to explain these results since longer reaction times with increasing disparity between the prime and target stimuli would be expected.

According to the neurophysiologic definition of top-down processing, which is defined as recurrent feedback from higher-level brain areas to lower-level brain areas no top-down but a lateral interference would thus cause the fastened recognition in both postulated mechanisms. The third possibility however, which is compatible with the neurophysiologic definition of top-down influence, is that the preceding object could 'pre-activate' the visual memory representations of the associated object in different views through backward connections from associative memory. This top-down 'pre-activation' then could cause the fastened activation of an associated object in a non-canonical view. The idea of high-level nodes that are activated by context is also proposed by Graboi and Lisman (2003). They argue in their hierarchical model that high level nodes are activated even before an item is presented, if the item is consistent with the current context. The different interpretations of the results illustrate that a concluding explanation for the fastened recognition of objects in non-canonical views with the use of top-down processing in a neurophysiologic sense remains outstanding. To investigate these questions further neuro-imaging studies (e.g., fMRI) could provide a deeper insight into the neuronal processes.

5.6 APPENDIX

Stimuli used in Experiment 1. can = canonical view, non-can = non-canonical view.

5.7 REFERENCES

Bar, M. (2003). A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition. *Journal of Cognitive Neuroscience, 15(4),* 600-609.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. Psychological Review, 94, 115-147.

Biederman, I., Blickle, T. W., Teitelbaum, R. C., Klatsky, G. J., & Mezzanotte, R. J. (1988). Object identification in nonscene displays. *Journal of Experimental Psychology: Human Learning, Memory, and Cognition, 14,* 456-467.

Biederman, I., Mezzanotte, R. J., & Rahinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14,* 143-177.

Blanz, V., Tarr, M.J., & Bülthoff, H. H., (1999). What object attributes determine canonical views? *Perception, 28,* 575-599.

Boyce, S.J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 15,* 556-566.

Bülthoff, H. H. & Edelman, S. (1992). Psychophysical support or a two-dimensional view interpolation theory of object recognition. *Proc. Natl Acad. Sci. USA 89,* 60-64.

Edelman, S. (1999). *Representation and recognition in vision.* Cambridge, MA: MIT Press.

Edelman, S. & Newell, F. N. (1998). *On the representation of object structure in human vision: evidence from differential priming of shape and location.* CSRP 500, University of Sussex.

Edelman, S., & H. H. Bülthoff (1992). Orientation dependence in the recognition of familiar and novel views of 3D objects, *Vision Research 32,* 2385-2400.

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society of London Series B Biological Sciences*, 269, 1939-1947.

Graf, M., Kaping, D., & Bülthoff, H. H. (2005). Orientation congruency effects for familiar objects: Coordinate transformations in object recognition. *Psychological Science, 16*, 214-221.

Gibson, B. S. & Peterson, M. A. (1994). Does orientation-independent object recognition precede orientation-dependent recognition? *Journal of Experimental Psychology: Human Perception and Performance, 20,* 299-316.

Graboi, D., & Lisman, J. (2003). Recognition by top-down and bottom-up processing in cortex: the control of selective attention. *Journal of Neurophysiology, 90*, 798-810.

Hayward, W. G. (2003). After the viewpoint debate: where next in object recognition? *Trends in Cognitive Sciences, 7,* 425-427.

Hayward, W. G., & Tarr, M. J. (1997). Testing conditions for viewpoint invariance in object recognition. *Journal of Experimental Psychology: Human Perception and Performance, 23,* 1511-1521.

Hayward, W. G., & Williams, P. (2000). Viewpoint dependence and object discriminability. *Psychological Science, 11,* 7-12.

Henderson, J. M., & Hollingworth, A. (1999). High-Level Scene Perception. *Annual Review of Psychology, 50,* 243-271.

Henderson, J.M., Pollatsek A., & Rayner, K. (1987). Effects of foveal priming and extrafoveal preview on object identification. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 449-463.

Hill, H., Schyns, P.G., & Akamatsu, S. (1997). Information and viewpoint dependence in face recognition. *Cognition, 62,* 201-222.

Hollingworth, A., & Henderson, J. M. (1998). Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General, 127,* 398-415.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review, 99,* 480-517.

Humphreys, G.W., Riddoch, M.J., & Price, C.J. (1997). Top-down processes in object identication: evidence from experimental psychology, neuropsychology and functional anatomy. *Phil. Trans. R. Soc. Lond. B, 352,* 1275 – 1282.

Jolicoeur, P. (1992). Orientation Congruency Effects in Visual search. *Canadian Journal of Psychology, 46,* 280-305.

Jolicoeur, P., & Humphrey, G.K. (1998). Perception of rotated two-dimensional and three-dimensional objects and visual shapes. In V. Walsh & J. Kulikowski (Eds). *Perceptual constancy: Why things look as they do* (pp. 69-123). Cambridge, U.K.:Cambridge University Press.

Knoblich, U., Riesenhuber, M., Freedman, D.J., Miller, E.K., & Poggio, T. (2002). Visual Categorization: How the Monkey Brain Does It. In *Biologically Motivated Computer Vision*, Second International Workshop, BMCV 2002, Tübingen, Germany.

Kosslyn, S. M. (1994). *Image and brain. The resolution of the imagery debate.* Cambridge, MA: MIT Press.

Lawson, R., & Humphreys, G. W. (1996). View-specificity in object processing: Evidence from picture matching. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 395-416.

Lawson, R., & Humphreys, G. W. (1998). View-specific effects of depth rotation and foreshortening on the initial recognition and priming of familiar objects. *Perception and Psychophysics, 60,* 1052-1066.

Liter, J.C. & Bülthoff, H.H. (1997). View canonicality affects naming but not name verification of common objects. *Technical Report No. 051,* Max-Planck Institute for Biological Cybernetics, Tuebingen.

Lorch, R.F., Balota, D.A., & Stamm, E.G. (1986). Locus of inhibition effects in the priming of lexical decisions: pre- or postlexical access? *Memory and Cognition, 14,* 95-103.

Marr, D. (1982). *Vision.* San Francisco: Freeman.

Mumford, D. (1992). On the computational architecture of the neocortex. II. The role of cortico-cortical loops. *Biological Cybernetics, 66*, 241-251.

Murray, J.E. (1997). Flipping and spinning. Spatial transformation procedures in the identification of rotated natural objects. *Memory & Cognition, 25,* 96-105.^

Murray, J.E. (1999). Orientation-specific effects in picture matching and naming. *Memory and Cognition, 27,* 878-889.

Neely, J.H. (1991). Semantic priming effects in visual word recognition: a selective review of current findings and theories. In D. Besner and G.W. Humphreys (Eds.), *Basic Processes in Reading* (pp. 264-336). Hillsdale: Laurence Erlbaum.

Neely, J.H., & Keefe, D.E., & Ross, K.L. (1989). Semantic priming in the lexical decision task roles of prospective prime-generated expectancies and retrospective semantic matching, *Jorunal of Experimental Psychology: Learning, Memory and Cognition, 15*, 1003-1019.

Newell, F.N. & Findlay, J.M. (1997). The Effect of Depth Rotation on Object Identification. *Perception, 26,* 1231-1257.

Palmer, S. E. (1975). The effects of contextual scenes on the identification of objects. *Memory & Cognition, 3,* 519-526.

Palmer, S. E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In J. Long & A. Baddeley (Eds.), *Attention and Performance,* IX (pp. 135 151). Hillsdale, N.J. Erlbaum.

Perrett, D. I., Oram, M. W., & Ashbridge, E. (1998). Evidence accumulation in cell populations responsive to faces: An account of generalisation of recognition without mental transformations. Cognition, 67, 111-145.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical Models of Object Recognition in Cortex. *Nature Neuroscience, 2,* 1019-1025.

Schwaninger, A. (2004a). Objekterkennung und Signaldetektion. In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 106-130). Bern: Huber.

Takano, Y. (1989). Perception of rotated forms: a theory of information types, *Cognitive Psychology, 21,* 1-59.

Tarr, M. J. (1995). Rotating objects to recognize them: A case study of the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review, 2,* 55-82.

Tarr, M. J., & Bülthoff, H. H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? Comment on Biederman and Gerhardstein (1993). *Journal of Experimental Psychology: Human Perception and Performance, 21,* 1494-505.

Tarr, M. J., & Bülthoff, H. H. (1998). Image-based object recognition in man, monkey and machine. In M.J. Tarr & H. Bülthoff (Eds.), *Object recognition in man, monkey, and machine* (pp. 1-20). Cambridge, MA: MIT Press.

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology, 21,* 233-282.

Ullman, S. (1995). Sequence-seeking and counter-streams: a model for information flow in the cortex. *Cerebral Cortex*, 5, 1-11.

Ullman, S. (1996). High-level vision: Object recognition and visual cognition (1st ed.). Cambridge, MA: MIT Press.

PART II
HUMAN FACTORS IN AVIATION SECURITY

6 MEASURING VISUAL ABILITIES AND VISUAL KNOWLEDGE OF AVIATION SECURITY SCREENERS

6.1 ABSTRACT

A central aspect of airport security is reliable detection of forbidden objects in passenger bags using X-ray screening equipment. Human recognition involves visual processing of the X-ray image and matching items with object representations stored in visual memory. Thus, without knowing which objects are forbidden and what they look like, prohibited items are difficult to recognize (aspect of visual knowledge). In order to measure whether a screener has acquired the necessary visual knowledge, we have applied the prohibited items test (PIT). This test contains different forbidden items according to international prohibited items lists. The items are placed in X-ray images of passenger bags so that the object shapes can be seen relatively well. Since all images can be inspected for 10 seconds, failing to recognize a threat item can be mainly attributed to a lack of visual knowledge.

The object recognition test (ORT) is more related to visual processing and encoding. Three image-based factors can be distinguished that challenge different visual processing abilities. First, depending on the rotation within a bag, an object can be more or less difficult to recognize (effect of viewpoint). Second, prohibited items can be more or less superimposed by other objects, which can impair detection performance (effect of superposition). Third, the number and type of other objects in a bag can challenge visual search and processing capacity (effect of bag complexity). The ORT has been developed to measure how well screeners can cope with these image-based factors. This test contains only guns and knives, placed into bags in different views with different superposition and complexity levels. Detection performance is determined by the ability of a screener to detect threat items despite rotation, superposition and bag complexity. Since the shapes of guns and knives are usually known well even by novices, the aspect of visual threat object knowledge is of minor importance in this test.

A total of 134 aviation security screeners and 134 novices participated in this study. Detection performance was measured using A'. The three image-based factors of the ORT were validated. The effect of view, superposition and bag complexity were highly significant. The validity of the PIT was examined by comparing the two participant groups. Large differences were found in detection performance between screeners and novices for the PIT. This result is consistent with the assumption that the PIT measures aspects related to visual knowledge. Although screeners were also better than novices

in the ORT, the relative difference was much smaller. This result is consistent with the assumption that the ORT measures image-based factors that are related to visual processing abilities whereas the PIT is more related to visual knowledge. For both tests, large inter-individual differences were found. Reliability was high for both participant groups and tests, indicating that they can be used for measuring performance on an individual basis. The application of the ORT and PIT for screener certification and competency assessment are discussed.

6.2 INTRODUCTION

The importance of aviation security has changed dramatically in the last years. As a consequence of the new threat situation large investments into technology have been made. State-of-the-art X-ray machines provide high resolution images, many image enhancement features and even automatic detection of explosive material. However, it is becoming clear since recently that the best technology is only as valuable as the humans that operate it. Indeed, reliable recognition of threat items in X-ray images of passenger bags is a demanding task. Consider the images depicted in Figure 1. Each of the two bags contains a threat item that could be used to severely harm people. Even though most people would probably recognize prohibited items like the gas spray in Figure 1a when depicted in a photograph, this and other threat objects are relatively hard to recognize for novices because the shape features look quite different in an X-ray image than in reality. Other dangerous items (e.g. the switchblade knife in Figure 1b) might be missed by a novice because they look similar to harmless objects (e.g. a pen). Several other threat objects are usually not encountered in real life (e.g. improvised explosive devices, IEDs), which stresses the

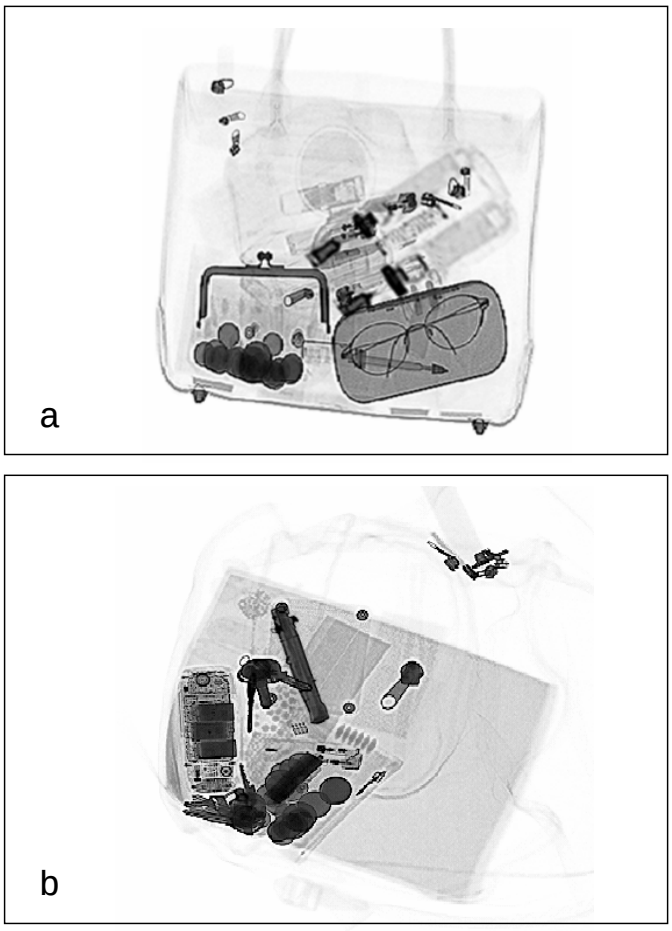

Figure 1. Examples of prohibited items in x-ray images of passenger bags. (a) gas spray in the centre of the baggage below the eyeglasses, (b) switchblade knife slightly above the centre of the baggage next to the keys.

importance of computer-based training in order to achieve a high detection performance within a few seconds of inspection time (Schwaninger & Hofer, 2004).

In short, the knowledge about which items are prohibited and what they look like in an X-ray image is certainly an important determinant for detection performance. The Prohibited Items Test (PIT) has been developed to measure this knowledge-based component and it therefore contains a large number of different forbidden objects according to international prohibited items lists (Schwaninger, 2004a).

As pointed out by Schwaninger (2003) several image-based effects influence how well threat items can be recognized in X-ray images (Figure 2). Viewpoint can strongly affect recognition performance, which has been shown previously in many object recognition studies (for reviews see Graf, Schwaninger & Bülthoff, 2002; Schwaninger, 2004b; Tarr & Bülthoff, 1995, 1999). Since objects are often superimposed on each other in X-ray images, the degree of superposition can affect detection performance substantially. Another image-based factor is bag complexity, which is determined by the type and number of objects in a bag.

The Object Recognition Test (ORT) has been developed to measure how well screeners can cope with such image-based factors (Schwaninger, 2004c). In order to reduce effects of visual knowledge, only guns and knives are used in this test, i.e. object shapes that are usually well known also by novices.

The purpose of this study is to investigate the role of image-based and knowledge based factors in X-ray screening using these two different tests.

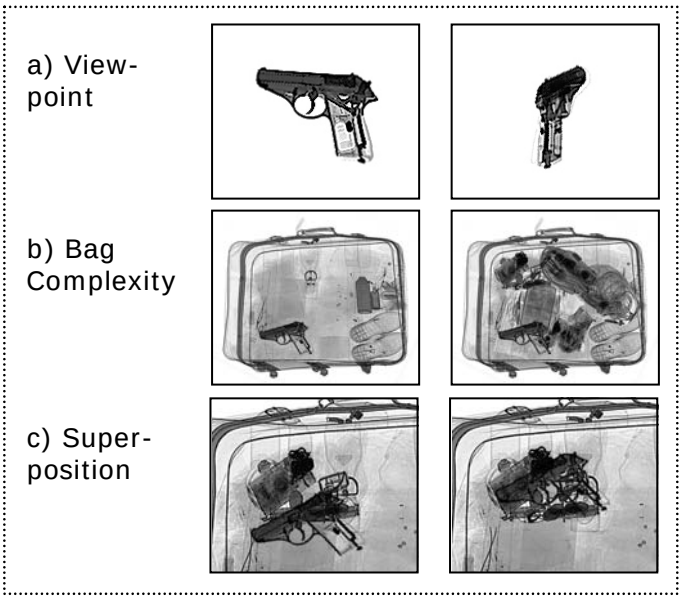

Figure 2. Image-based factors: (a) viewpoint of the threat item (canonical vs. non-canonical), (b) bag complexity (low vs. high), (c) superposition of the threat item (low vs. high).

To what extent screeners know which items are prohibited and what they look like in passenger bags is measured by the PIT. It includes prohibited items of different categories in X-ray images of passenger bags while keeping effects of view, superposition and bag complexity relatively constant. The objects are displayed in an easy view with a moderate degree of superposition in bags of limited complexity during 10 seconds per image. If a participant fails to detect a threat item it is therefore rather related

to a lack of visual knowledge than to an attention failure or visual processing capacity limitations. Since many different prohibited items with shapes that are often not known from everyday experience are used in the PIT, a substantial difference in detection performance between novices and screeners could be expected. The ORT measures how good someone can cope with image-based factors such as view, superposition and bag complexity. As mentioned above, only guns and knives are used in this test, i.e. object shapes that are well known by both screeners and novices. Therefore, smaller differences between screeners and novices might be expected for the ORT compared to the PIT. However, expertise might increase visual abilities that are necessary in order to cope with image difficulty resulting from effects of viewpoint, superposition and bag complexity. Therefore, the effect size of the interaction between image-based effects and expertise is an important measure in this study as well.

6.3 METHOD

6.3.1 PARTICIPANTS

A total of 268 participants took part in this study. Half of them were aviation security screeners, the other half were novices. All participants were tested with the ORT and then the PIT. The screener group consisted of 67 females and 67 males at the age between 24 and 57 years (M = 41.05 years, SD = 7.84 years). All of them had undergone initial class-room and on the job training and they had at least two years of work experience in airport security screening of carry-on bags.

The novices group consisted of 134 males between 21 and 26 years (M = 23.24 years, SD = 1.22 years).

6.3.2 MATERIALS AND PROCEDURE

6.3.2.1 PROHIBITED ITEMS TEST (PIT)

This test contains a wide spectrum of prohibited items which can be classified into seven categories according to international prohibited items lists (Schwaninger, 2004a). The PIT version used in this study included a total of 19 guns, 27 sharp objects, 14 blunt and hunt instruments, 5 highly inflammable substances, 17 explosives, 3 chemicals and 13 other prohibited items (e.g. buckshot, ivory). All prohibited items were depicted from an easy viewpoint and combined with a bag of medium complexity and low superposition, so that their shapes could be seen relatively well and the influence of image-based factors could be minimized. X-ray images were taken from Heimann 6040i machines and displayed in color. 68 bags contained one threat item, 6 bags contained two threat items, and 6 bags

contained three threat items. Each bag was shown twice resulting in a total of 160 trials. There were four blocks of 40 trials. Block order was counterbalanced across four groups of participants using a Latin Square design. Trial order was randomized within each block. Only responses to images containing one threat item were used for statistical analyses.

The PIT is fully computer-based and starts with a self-explanatory instruction, followed by a brief training session with eight examples to familiarize the participants with the procedure. Feedback is provided after each trial only in the introductory phase. Each X-ray image was displayed for a maximum of 10 seconds in the introductory and test phase. This duration is long enough to ensure that missing a threat item can be mainly attributed to a lack of visual knowledge rather than a failure of attention. For each image, participants had to decide whether the bag was OK (no threat) or NOT OK (threat) and indicate on a slider how sure they were in their decision (confidence ratings on a 50 point scale). In addition, participants had to indicate the threat category of the prohibited item(s) by clicking the corresponding buttons on the screen (for NOT OK decisions only). Pressing the space bar displayed the next image. As the test was subdivided into four blocks, participants were allowed to take a short break after a block was completed.

6.3.2.2 OBJECT RECOGNITION TEST (ORT)

As explained in the introduction, Schwaninger (2003) pointed out that image-based factors such as viewpoint, superposition and bag complexity can substantially affect detection performance in X-ray images. The ORT has been designed to measure how well people can cope with such image-based factors rather than measuring knowledge-based determinants of threat detection performance (Schwaninger, 2004c). To this end, guns and knives with the blade open are used in the ORT, i.e. object shapes that can be assumed to be known by most people. In addition, all guns and knives are shown for 10 seconds before the test starts, which further reduces the role of knowledge based factors in this test.

In reality, a threat object can be depicted from a difficult viewpoint in a close-packed bag and be superimposed by other objects. The X-ray images used in the ORT vary systematically in image difficulty by varying the degree of view difficulty, bag complexity and superposition, both independently and in combination. This makes it possible to investigate main effects as well as interactions between the image-based factors. All X-ray images of the ORT are in black-and-white, as color per se is mainly diagnostic for the material of objects in the bag and thus could be primarily helpful for experts.

Eight guns and eight knives with common shapes were used. Each gun and each knife was displayed in an easy view and a rotated view to measure the effect of viewpoint. In order to equalize image difficulty resulting from viewpoint changes, guns were more rotated than knives based on results of a pilot study. Each view was combined with two bags of low complexity, once with low superposition and once with high superposition. These combinations were also generated using two closed-packed bags with a higher degree of bag complexity. In addition, each bag was presented once with and once without the threat item. Thus, there were a total of 256 trials: 2 weapons (guns, knives) * 8 (exemplars) * 2 (views) * 2 (bag complexities) * 2 (superpositions) * 2 (harmless vs. threat images). There were four blocks of 64 trials each. The order of blocks was counterbalanced across four groups of participants using a Latin Square. Within each block the order of trials was random.

The ORT is fully computer-based. After task instructions an introductory session followed using 2 guns and 2 knives not displayed in the test phase. Feedback was provided after each trial but only in the introductory phase. Prior to the test phase, the eight guns and eight knives used at test were presented for 10 seconds, respectively. Half of the guns and knives were shown in an easy view and half of them were depicted in a rotated view. At test, each object was presented in the easy and the rotated view with low and high superposition and with low and high bag complexity. Each image was displayed for 4 seconds. This duration was chosen to match the demands of high passenger flow where average X-ray image inspection time at checkpoints is in the range of 3-5 seconds. For each X-ray image, participants had to decide whether the X-ray images contained one of the guns or knives shown in the introductory phase or not (NOT OK or OK response). Confidence ratings had to be provided by changing the position of a slider (90 point scale). The next trial was started by pressing the space bar. Short breaks were possible after completing one of the four blocks.

6.4 RESULTS

It is important to take the hit rate as well as the false alarm rate into account if threat and non-threat images are used in a computer-based test requiring OK and NOT OK responses. The reason is simple: A candidate could achieve a hit rate of 100 percent simply by judging all bags as being NOT OK. Whether a high hit rate reflects good visual detection performance, or just a lenient response bias, can only be determined if the false alarm rate is considered too. Psychophysics provides several methods in order to derive more valid estimates based on hit and false alarm rates. A

well-known measure from signal detection theory is d' (Green & Swets, 1966). It equals z(H) – z(FA) whereas H denotes the hit rate, FA the false alarm rate and z represents the transformation into z-scores (standard deviation units). An often used "non-parametric" measure is A' (Pollack & Norman, 1964). This measure represents an estimate of the area under an ROC curve that is specified by only one data point. More specifically, A' corresponds to the average area for the two linear ROC curves that maximize and minimize the hit rate. The term "non-parametric" is a bit misleading because it only refers to the fact that the computation of A' doesn't require an priori assumption about the underlying distributions (Pastore et al., 2003; MacMillan & Creelman, 1991). This has sometimes been regarded as an advantage over SDT measures such as d' and Δm (for a more detailed discussion of this issue see also Hofer & Schwaninger, 2004). Although only A' data are reported in this study, it should be stressed that similar results were obtained for d' data. Moreover, correlations between A' and d' were very high for both tests and screeners groups (ORT: r = .94 for screeners and r = .97 for novices, PIT: r = .95 for screeners and r = .98 for novices, all $p < .001$).

The results section is organized as follows. First, ANOVA results of the ORT are presented. These analyses were conducted to investigate whether detection performance of aviation security screeners and novices is affected by image-based factors. In addition, the effect of expertise on the three image-based factors measured by the ORT was examined. Second, overall detection performance in the ORT is compared to overall detection performance in the PIT[1]. More specifically, the effect of expertise on image-based factors and knowledge-based factors is analyzed, comparing detection performance of aviations security screeners with that of novices in the two tests. Finally, the results of reliability analyses are presented which were conducted to evaluate whether the ORT and PIT can be used for measuring detection performance on an individual basis.

6.4.1 ORT AND ABILITIES TO COPE WITH IMAGE-BASED FACTORS

A' scores calculated from hit and false alarm rates of the ORT were subjected to three-way analyses of variance (ANOVA) with the three within-participants factors view, bag complexity and superposition. Results of aviation security screeners show that there were significant main effects of view (easy vs. rotated) with an effect size of $\eta^2 = .71$, $F(1, 133) = 318.59$, $MSE = 0.003$, $p < .001$, bag complexity (low vs. high) $\eta^2 = .83$, $F(1, 133) =$

[1] A' scores for the PIT were calculated using the responses to images of the following categories: guns, sharp objects, hunt and blunt instruments.

652.96, MSE = 0.003, p < .001, and superposition (low vs. high) η^2 = .61, F(1, 133) = 203.73, MSE = 0.003, p < .001. The following two-way interactions were significant: View * superposition η^2 = .12, F(1, 133) = 17.91, MSE = 0.002, p < .001, bag complexity * superposition η^2 = .12, F(1, 133) = 18.22, MSE = 0.002, p < .001. Note however, that the effect sizes of these interactions are rather low when compared to the effect sizes of the main effects. All other interactions were not significant. In short, there were clear main effects of view, bag complexity and superposition with very large effect sizes (see also conventions by Cohen, 1988). Some interactions reached statistical significance, but the effect sizes were relatively small when compared to the effect sizes of the main effects.

Similar results could be observed for novices. Again, there were significant main effects of view (easy vs. rotated) η^2 = .76, F(1, 133) = 428.33, MSE = 0.005, p < .001, bag complexity (low vs. high) η^2 = .72, F(1, 133) = 333.14, MSE = 0.005, p < .001, and superposition (low vs. high) η^2 = .63, F(1, 133) = 228.09, MSE = 0.004, p < .001. All two-way interactions were significant:

View * bag complexity η^2 = .06, F(1, 133) = 9.07, MSE = 0.004, p < .01, view * superposition η^2 = .07, F(1, 133) = 10.43, MSE = 0.004, p < .01, bag complexity * superposition η^2 = .15, F(1, 133) = 23.15, MSE = 0.004, p < .001. The three-way interaction between view, bag complexity and superposition also reached statistical significance, η^2 = .03, F(1, 133) = 4.14, MSE = 0.004, p < .05. As for screeners, very large effect sizes were found for main effects whereas the interactions showed much smaller effect sizes.

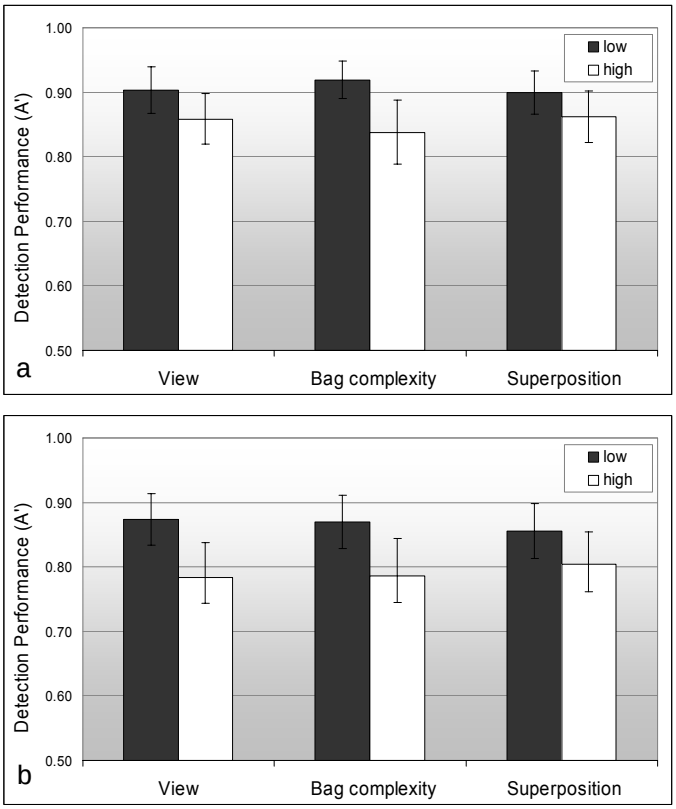

Figure 3. Detection performance (A') in the ORT with standard deviations: (a) for aviations security screeners, (b) for novices.

Figure 3 shows the main effects of each of the three image-based factors, averaged across the other two factors.

A comparison of Figure 3a (aviation security screeners) and Figure 3b (novices) reveals that screeners were slightly better than novices while both screener groups are substantially affected by the image-based factors view, bag complexity, and superposition. In order to examine whether expertise has a differential effect

on these image-based factors, a four-way analysis of variance (ANOVA) with the within-participants factors view, bag complexity and superposition and the between-participant factor expertise was computed. There were again significant main effects of view (easy vs. rotated) $\eta^2 = .74$, $F(1, 266) = 744.57$, $MSE = 0.004$, $p < .001$, bag complexity (low vs. high) $\eta^2 = .77$, $F(1, 266) = 884.75$, $MSE = 0.004$, $p < .001$, and superposition (low vs. high) $\eta^2 = .62$, $F(1, 266) = 428.20$, $MSE = 0.004$, $p < .001$. Two-way interactions between view and bag complexity $\eta^2 = .04$, $F(1, 266) = 10.23$, $MSE = 0.003$, $p < .01$, view and superposition $\eta^2 = .09$, $F(1, 266) = 26.17$, $MSE = 0.003$, $p < .001$, view and expertise $\eta^2 = .10$, $F(1, 266) = 30.52$, $p < .001$, and superposition and expertise $\eta^2 = .03$, $F(1, 266) = 9.39$, $p < .01$ were significant, as well as the tree-way interactions between view, bag complexity and superposition $\eta^2 = .02$, $F(1, 266) = 5.47$, $MSE = 0.003$, $p < .05$, and bag complexity, superposition and expertise $\eta^2 = .13$, $F(1, 266) = 41.13$, $p < .001$. Although these interactions were significant, all of them have relatively low effect sizes when compared to the main effects. All other interactions were not significant.

In short, these results indicate that the effects of image-based factors are apparent for novices as well as for aviation security screeners and expertise does only slightly reduce these effects of view, bag complexity and superposition.

6.4.2 PIT, VISUAL KNOWLEDGE AND EXPERTISE

In contrast to the ORT, the PIT has been developed to measure whether screeners know which items are prohibited and how they look like in X-ray images of passenger bags (Schwaninger, 2004a). Whereas in the ORT only guns and knives are used – object shapes that are also familiar to novices – the PIT contains all kinds of forbidden objects based on international prohibited items lists. In this test, all target objects are shown in an easy viewpoint with a moderate degree of superposition in bags of moderate bag complexity. As mentioned above, each image is shown for 10 seconds and therefore missing a threat item in the PIT can rather be attributed to a lack of visual knowledge than to an attention failure or visual processing capacity limitations. If detection performance in the PIT is indeed mainly determined by visual experience and training with X-ray images, large differences between novices and aviation security screeners should be observed in this test. As reported in the previous section, only moderate differences between novices and screeners were found for the ORT.

In order to compare relative difference between experts and novices for the PIT and ORT, overall hit and false alarm rates were used to compute

relative detection performance difference separately for the ORT and PIT using the following formula: $(A'_{experts} - A'_{novices}) / A'_{novices}$. Relative detection performance difference between experts and novices was indeed much higher for the PIT than for the ORT (15.89 percent vs. 6.05 percent). This is consistent with the view that the PIT measures visual knowledge dependent on training and expertise, whereas the ORT measures more stable visual abilities used to cope with image-based factors such as effects of view, bag complexity and superposition.

This main finding was further analyzed using a two-way analysis of variance (ANOVA) with the within-participant factor test type (ORT, PIT) and the between-participant factor expertise using overall A' scores from each test. There was a significant effect of test type (ORT vs. PIT) $\eta^2 = .80$, $F(1, 266) = 1075.10$, $MSE = 0.002$, $p < .001$, a significant effect of expertise (experts vs. novices) $\eta^2 = .44$, $F(1, 266) = 206.11$, $MSE = 0.004$, $p < .001$, and a significant interaction of test type and expertise $\eta^2 = .20$, $F(1, 266) = 65.30$, $p < .001$. The interaction between test type and expertise is consistent with the hypothesis that the ORT measures rather image-based factors whereas the PIT measures rather knowledge-based factors.

It must also be noted however, that correlation analyses showed that the two tests are far from being orthogonal. Overall detection performance A' of the two tests correlates with $r = .51$, $p < .001$ for experts, and $r = .42$, $p < .001$ for novices. This could at least indicate that detection performance in PIT is not only determined by visual knowledge but also by visual abilities used to cope with image-based factors as measured by the ORT.

One potential argument against the analyses of this section could be that the expert group consisted of males and females, whereas the novices group consisted only of males. However, it is unlikely that gender effects can explain the differences found between experts and novices since no significant differences were found between male and female screeners, neither for the ORT ($p = .70$) nor for the PIT ($p = .78$).

6.4.3 RELIABILITY ANALYSES

Internal reliability was analyzed using Cronbach's Alpha and Guttman split-half coefficients separately for both participant groups (aviation security screeners and novices).

TABLE 1
RELIABILITY ANALYSES

Reliability Coefficients			PC SN	PC N	CR SN	CR N
PIT	Screeners	Cronbach's Alpha	.840	.878	.887	.924
		Split-half	.811	.915	.859	.948
	Novices	Cronbach's Alpha	.871	.877	.885	.914
		Split-half	.882	.862	.883	.890
ORT	Screeners	Cronbach's Alpha	.862	.934	.902	.962
		Split-half	.733	.813	.792	.887
	Novices	Cronbach's Alpha	.899	.910	.916	.959
		Split-half	.778	.810	.759	.907

Note. Cronbach's Alpha values and split-half reliabilities (Guttman) for both tests in each group (experts and novices separately) calculated for percentage correct (PC) and confidence ratings (CR) separately for signal plus noise trials (SN) and noise trials (N).

Analyses were computed for signal plus noise trials (bags including a threat item) and noise trials (harmless bags), respectively.

Reliability coefficients were computed on the basis of the percentage correct measures (i.e. hit and correct rejections), as well as on the basis of the screeners' confidence ratings (CR) for hits and correct rejections.

As can be seen in Table 1 high reliability coefficients were found for both tests and participant groups.

The results of section 6.4.1 have clearly shown that item difficulty in the ORT depends on the main effects and interactions between view, bag complexity and superposition. Therefore, the high internal consistency also found for the ORT is a nice example for the fact that a test can be homogenous and multifactorial (see also for example Kline, 2000).

6.5 DISCUSSION

The objective of this study was to examine the role of image-based and knowledge-based factors for detecting threat items in passenger bags. As pointed out by Schwaninger (2003), image-based factors such as effects of viewpoint, bag complexity and superposition can substantially affect detection performance. The ORT has been developed to measure how good a participant can cope with these image-based factors (Schwaninger, 2004c). This test contains guns and knives depicted in an easy and difficult view shown in bags with low and high bag complexity while being strongly or little superimposed by other objects. Main effects with large effect sizes were found for aviation security screeners as well as novices. While screeners

achieved a moderately better detection performance in the ORT, they were still significantly affected when threat items were rotated, superimposed by other objects or shown in complex bags. This result is consistent with the view that the ORT does measure visual abilities necessary to cope with image difficulty resulting from effects of viewpoint, bag complexity and superposition. Large inter-individual differences were found both for novices as well as experts. Internal reliability was very high for both groups. Therefore, this test could be a useful tool both for competency assessment of screeners as well as for pre-employment assessment purposes.

The PIT has been developed to measure whether a screener knows which items are prohibited and what they look like in X-ray images of passenger bags (Schwaninger, 2004a). In this test, all objects are depicted in an easy view. Bag complexity and superposition are moderate so that the threat item shapes are visible. Images are shown for 10 seconds, i.e. missing a threat item can be attributed to a lack of visual knowledge rather than to an attention failure or a visual processing capacity limitation. If the PIT is indeed related to visual knowledge based on expertise and training, large differences between novices and experts should be observed. Indeed, relative detection performance difference between novices and experts was about three times higher for the PIT than for the ORT. This result is consistent with the view that the PIT measures rather knowledge-based factors whereas the ORT measures rather visual abilities used for coping with image-based factors. As for the ORT, excellent reliability coefficients were found for the PIT. This test could therefore provide a useful tool for certification, competency and risk assessment as well as for quality control in general.

In summary, the results of this study confirm that X-ray detection performance relies on visual abilities necessary for coping with image-based effects such as view, bag complexity and superposition. Visual experience and training are necessary to know which items are prohibited and what they look like in X-ray images of passenger bags. Both aspects are prerequisites for a good screener and can be evaluated using the ORT and PIT.

6.6 REFERENCES

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum, Hillsdale.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) program,* Cognitive Vision Systems - CogVis (IST-2000-29375).

Green, D. M., & Swets, A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

Schwaninger, A. (2003). Detection systems: Screener evaluation and selection, *AIRPORT, 02*, 14-15.

Kline, P. (2000). *The Handbook of Psychological Testing (2nd edition)*, London: Routledge, 2000.

MacMillan, N. A., & and Creelman, C. D. (1991). Detection theory: A user's guide. Cambridge: University Press.

Pastore, R. E., Crawley, E. J., Berens, M.S., & Skelly, M.A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556-569.

Pollack, I., & Norman, D.A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science, 1*, 125-126.

Schwaninger, A., (2004a). *Prohibited items test (PIT): Test manual and user's guide.* Zurich: APSS.

Schwaninger, A. (2004b). Objekterkennung und Signaldetektion. In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 106-130). Bern: Huber.

Schwaninger, A., (2004c). *Object recognition test (ORT): Test manual and user's guide.* Zurich: APSS.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex : WIT Press.

Tarr, M. J., & Bülthoff, H.H. (1995). Is human object recognition better described by geon structural descriptions or by multiple views? J*ournal of Experimental Psychology: Human Perception and Performance, 21(6)*, 1494-1505.

Tarr, M. J., & Bülthoff, H.H. (1999). *Object recognition in man, monkey and machine*. Cambridge, Massachusetts: MIT Press.

7 THE X-RAY OBJECT RECOGNITION TEST (X-RAY ORT) – A RELIABLE AND VALID INSTRUMENT FOR MEASURING VISUAL ABILITIES NEEDED IN X-RAY SCREENING

7.1 ABSTRACT

Aviation security screening has become very important in recent years. It was shown in [1] that certain image-based factors influence detection when visually inspecting x-ray images of passenger bags. Threat items are more difficult to recognize when placed in close-packed bags (effect of bag complexity), when superimposed by other objects (effect of superposition), and when rotated (effect of viewpoint). The X-Ray Object Recognition Test (X-Ray ORT) was developed to measure the abilities needed to cope with these factors. In this study, we examined the reliability and validity of the X-Ray ORT based on a sample of 453 aviation security screeners and 453 novices. Cronbach's Alpha and split-half analysis revealed high reliability. Validity was examined using internal, convergent, discriminant and criterion-related validity estimates. The results show that the X-Ray ORT is a reliable and valid instrument for measuring visual abilities needed in x-ray screening. This makes the X-Ray ORT an interesting tool for competency and pre-employment assessment purposes.

7.2 INTRODUCTION

One of the most important tasks in airport security screening is the visual inspection of passenger bags using x-ray imaging systems. During rush hours, screeners have only a few seconds to decide whether a bag is OK (i.e. it contains no prohibited item) or NOT OK (i.e. it contains a prohibited item). Understanding the underlying visual cognition processes of this task is very important in order to train and select people appropriately for the x-ray screening job. A screener has to know which items are prohibited and what they look like in x-ray images of passenger bags. This is dependent on training and expertise [2, 3]. In addition to such knowledge-based factors, [1] and [4] have identified three image-based factors, which are illustrated in Figure 1. Threat items are more difficult to detect when they are in a close-packed bag (effect of bag complexity). Objects in x-ray images are often superimposed by other objects, which can also affect detection performance (effect of superposition). When threat objects are rotated they can become more difficult to recognize (effect of viewpoint).

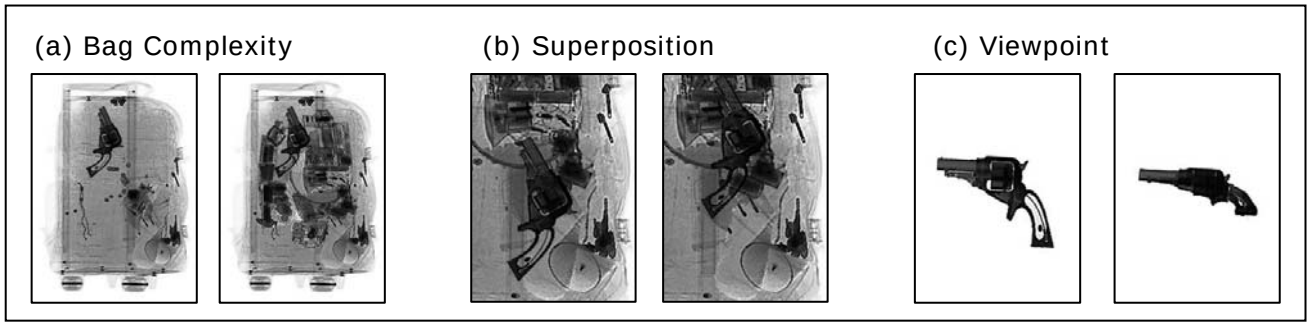

Figure 1: Image-based factors relevant in x-ray screening: (a) bag complexity, (b) superposition, (c) viewpoint.

The X-Ray Object Recognition Test (X-Ray ORT) is a tool to measure the visual abilities needed to cope with these image-based factors [1, 4]. In this study we examined the reliability and validity of the X-Ray ORT. Reliability measures, such as Cronbach's Alpha and split-half reliabilities were assessed with two groups (novices and experts) of 453 participants each. Validity estimates included internal, convergent, discriminant and criterion-related measures.

7.3 METHOD

7.3.1 PARTICIPANTS

453 aviation security screeners (141 male and 312 female) between 24 and 65 years (M = 48.94 years, SD = 9.09 years) and 453 novices (333 male and 120 female) between 19 and 56 years (M = 36.44 years, SD = 10.77 years) participated in this study. All screeners had at least three years of experience in x-ray screening.

7.3.2 MATERIALS AND PROCEDURE

In the X-Ray ORT, x-ray images of passenger bags are shown in black and white only because novices do not know how to interpret color information (which is in fact used to code different materials). To further reduce knowledge-based factors resulting from training or experience, only guns and knives with common shapes are used in the X-Ray ORT. Moreover, all threat items are presented before the test starts (8 guns for ten seconds followed by 8 knives for 10 seconds). Half of the threat items are shown in a frontal view, the other half in a rotated view.

All threat items are combined with bags of different bag complexities (low and high) using different levels of superposition (low and high). Each threat item is shown from two viewpoints (easy vs. difficult). The difficulty levels of bag complexity, superposition and viewpoint were determined visually by two raters. Each bag was used twice, once with a threat item (threat image) and once without (harmless image). Thus, the X-Ray ORT consists of a total of

256 test trials: 2 threat categories (guns, knives) * 8 (exemplars) * 2 (bag complexities) * 2 (superpositions) * 2 (views) * 2 (harmless images vs. threat images). Based on results from pilot studies, guns were more superimposed by other items in the bag and more rotated than knives in order to achieve a similar level of image difficulty.

The task in the X-Ray ORT is to visually inspect x-ray images of passenger bags for the presence of a gun or a knife. Each image is presented for 4 seconds on the screen in order to match visual inspection times at airports during periods of high passenger flow. For each trial, test candidates have to decide whether the bag is OK (no threat item included) or NOT OK (gun or a knife included) and indicate on a 90 point rating scale how sure they are in their decision (confidence ratings). All responses are made by clicking buttons on the screen. By pressing the space bar, the next trial is initiated.

Before the actual test starts, candidates are exposed to several screens with instructions as well as eight practice trials (half of them with a threat item and half of them without). None of the threat items and bags of the practice trials are used in the actual test. Whereas practice trials contain feedback on whether the correct response was given (OK vs. NOT OK), the actual test does not contain any feedback. The test is subdivided into four blocks and participants are allowed to take a short break after finishing a block. Trials are randomized within each block and block order is counterbalanced across four groups of participants using a Latin square design. The X-Ray ORT takes about 45 minutes to complete.

7.4 RESULTS

7.4.1 RELIABILITY OF THE X-RAY ORT

Cronbach's Alpha and Guttman split-half reliabilities were calculated for novices and experts. Reliability measures were based on hits and correct rejections (PC = percentage correct) as well as on confidence ratings (CR). Reliability was calculated separately for x-ray images of bags including a threat item (SN trials) and for x-ray images of harmless bags (N trials). The high reliability coefficients in Table 1 show that the X-Ray ORT is a reliable instrument for measuring visual abilities that are needed when visually inspecting x-ray images of passenger bags.

RELIABILITY ANALYSES

Reliability Coefficients			PC SN	PC N	CR SN	CR N
ORT	Screeners	Alpha	.887	.944	.926	.966
		Split-half	.781	.840	.840	.904
	Novices	Alpha	.907	.946	.932	.970
		Split-half	.778	.871	.807	.939

Table 1. PC = Percent Correct, CR = Confidence Ratings, SN = Bags containing a threat ("Signal plus Noise Trials"), N = Bags containing no threat ("Noise-Trials").

7.4.2 VALIDITY OF THE X-RAY ORT

Individual A' scores were calculated based on the percentage of hits and false alarms over all trials of the X-Ray ORT for each participant. The advantage of A' over d' is that it requires no a priori assumption about the underlying noise and signal plus noise distributions. For further information on these and other detection measures see [5, 6, 7].

7.4.2.1 INTERNAL VALIDITY

Individual A' scores were subjected to a three-way analysis of variance (ANOVA) with bag complexity, superposition and view difficulty as within-participant factors. This analysis was done for both groups of participants (experts and novices) separately. The main effects are displayed in Figure 2. ANOVA results of aviation security screeners showed highly significant main effects of bag complexity (low vs. high) with an effect size of $\eta^2 = .80$, $F(1, 452) = 1851.83$, $p < .001$, superposition (low vs. high) $\eta^2 = .55$, $F(1, 452) = 548.10$, $p < .001$, and view (easy vs. difficult) $\eta^2 = .70$, $F(1, 452) = 1044.01$, $p < .001$. Some interactions reached statistical significance but their effect sizes η^2 were small relative to the effect sizes of the main effects: bag complexity * superposition $\eta^2 = .06$, $F(1, 452) = 27.69$, $p < .001$, superposition * view $\eta^2 = .08$, $F(1, 452) = 37.90$, $p < .001$, and bag complexity * superposition * view $\eta^2 = .01$, $F(1, 452) = 6.55$, $p < .05$. Similar results were observed for novices. There were again highly significant main effects with large effect sizes: bag complexity (low vs. high) $\eta^2 = .69$, $F(1, 452) = 1012.20$, $p < .001$, superposition (low vs. high) $\eta^2 = .64$, $F(1, 452) = 817.19$, $p < .001$, and view (easy vs. difficult) $\eta^2 = .72$, $F(1, 452) = 1137.67$, $p < .001$. Again, some interactions were significant, but their effect sizes were rather small when compared to the effect sizes of the main effects. bag complexity * superposition $\eta^2 = .10$, $F(1, 452) = 48.01$, $p < .001$, bag complexity * view $\eta^2 = .10$, $F(1, 452) = 51.25$, $p < .001$, superposition * view $\eta^2 = .11$, $F(1, 452) = 55.35$, $p < .001$ and bag complexity * superposition * view $\eta 2 = .02$, $F(1, 452) = 8.64$, $p < .01$.

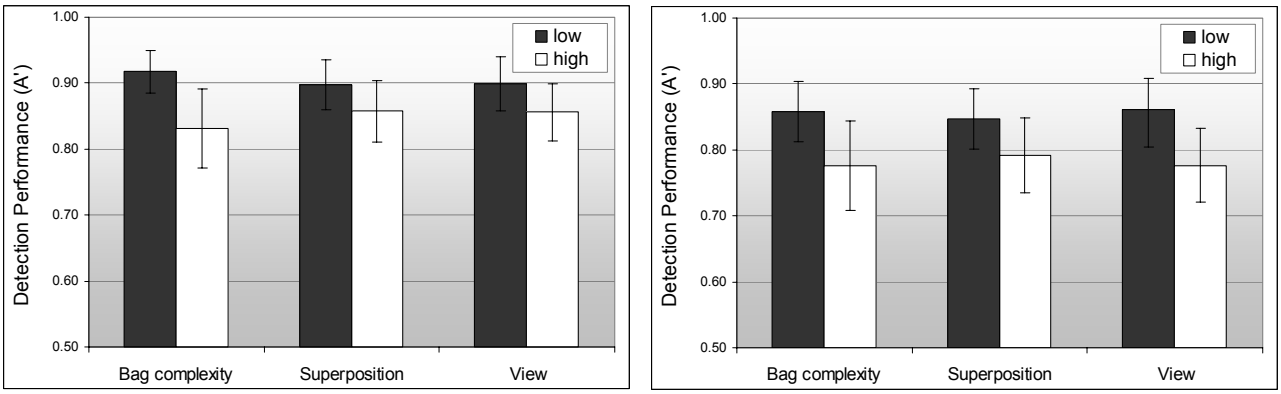

Figure 2: Effects of image-based factors in the X-Ray ORT, error bars represent standard deviations: LEFT: aviation security screeners, RIGHT: novices.

In summary, large main effects of image-based factors (bag complexity, superposition, and view difficulty) were found both for novices and experts. The large variances between individuals (see standard deviations in Figure 2) show that people differ remarkably with regard to how well they can cope with image difficulty resulting from these image-based effects. Interestingly, only small mean differences in A' between novices and experts were found. This is consistent with the assumption that the X-Ray ORT measures relatively stable visual abilities that are needed to cope with effects of bag complexity, superposition, and view difficulty.

However, we are currently conducting further studies in order to investigate whether these abilities can be trained when using an individually adaptive computer-based training system (X-Ray Tutor) that takes the image-based effects into account.

7.4.2.2 CONVERGENT AND DISCRIMINANT VALIDITY

Convergent validity was examined using the X-Ray ORT data from the aviation security screener group since all of them have taken also the Prohibited Items Test (PIT). The PIT is an image interpretation competency test using color x-ray images that contain different kinds of forbidden objects according to international prohibited items lists (for details see [1]). A' scores in the ORT correlated significantly with A' scores in the PIT, $r = .61$, $p < .001$, indicating high convergent validity. Discriminant validity was tested by correlating the X-Ray ORT with results obtained with the CBQ. The CBQ is a computer based multiple choice questionnaire about safety and security regulations on airports. As expected, the correlation with the X-Ray ORT was rather low, $r = .27$, indicating sufficient discriminant validity.

7.4.2.3 CRITERION-RELATED VALIDITY

Criterion-related validity was examined by correlating X-Ray ORT scores with on the job performance measured with threat image projection (TIP).

With this technology of current x-ray screening equipment it is possible to display fictional threat images during regular x-ray screening operations. Screeners receive feedback after each TIP image so that no negative impact on the screening operation occurs. The TIP library used in this study consisted of 1028 combined threat images (CTIs). These CTIs were created as follows: 64 improvised explosive devices (IEDs) were combined with 8 bags of different image difficulties rated by 8 x-ray screening experts. Each bag was also displayed without the IED. A TIP to bag ratio of 1 to 30 was used. This means that one TIP image was shown within about 30 x-ray images of real passenger bags. Half of the TIP images contained a threat item, the other half did not. This allowed obtaining valid hit and false alarm rates (see [6] for further information). TIP data was available from 86 screeners. On the job performance was estimated using TIP data aggregated over 17 months. A' and d' scores were calculated from hit and false alarm rates in TIP and in the X-Ray ORT. Large correlations between X-Ray ORT and TIP performance were found: $r = .41$, $p < .001$ for A' scores and $r = .51$, $p < .001$ for d' scores. These rather high correlations suggest that the abilities measured by the X-Ray ORT are indeed very important determinants of on the job performance in x-ray screening.

7.5 DISCUSSION

According to [1] and [4] detection of threat items in x-ray images of passenger bags depends on image-based factors such as bag complexity, superposition by other objects, and view difficulty of the threat item resulting from its position within the bag. The X-Ray ORT has been developed to measure how well people can cope with these image-based factors. In this study, the reliability and validity of the X-Ray ORT was examined. Cronbach's Alpha and split half analyses revealed that this test is a very reliable instrument. Cronbach's Alpha coefficients were found to be high in both samples of participants ($a > .89$ for experts and $a > .91$ for novices). Further evidence of reliability was revealed by split-half reliabilities (Guttman) which were quite high as well ($r > .78$ for both groups). Internal validity was examined using ANOVA. Highly significant main effects with large effect sizes were found for bag complexity, superposition, and view difficulty. Whereas some interactions reached statistical significance, their effect sizes were rather small when compared to the main effects. This indicates high internal validity regarding the assumption of three image-based factors that are conceptually independent. It should also be noted that large differences between individuals were found with regard to how well they could cope with effects of bag complexity, superposition, and view difficulty. Interestingly,

this accounted both for novices and screeners. Convergent validity was assessed by correlating X-Ray ORT scores with the results in the PIT, which is a computer-based image interpretation competency test. The large correlation of r = .61 supported convergent validity. Discriminant validity was estimated by correlating with the CBQ, a computer-based multiple choice exam on theoretical knowledge needed in airport security operations. In order to support discriminant validity a low correlation should be found. This was the indeed case since the X-Ray ORT correlated with CBQ scores only with r = .24. Criterion-related validity was calculated by correlating detection scores in the X-Ray ORT with on the job performance measured using threat image projection (TIP). Correlations of r = .41 using A' scores and r = .51 using d' scores indicated good criterion-related validity.

In summary, the results of different reliability and validity analyses showed that this test provides a very useful, reliable and valid instrument to assess visual abilities needed in x-ray screening of passenger bags. This makes the X-Ray ORT an interesting tool for competency and pre-employment assessment purposes in airport security and other areas in which x-ray screening is applied.

7.6 REFERENCES

Green, D.M., & Swets, J.A. (1966). *Signal detection theory and psychophysics,* New York: Wiley.

Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

MacMillan, N.A., & Creelman, C.D. (1991). *Detection theory: A user's guide.* Cambridge: University Press.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.

Schwaninger, A., & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex: WIT Press.

Schwaninger, A. (2003). Evaluation and selection of airport security screeners, *AIRPORT, 02/2003,* 14-15.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International , FEB/2004,* 31-36.

8 RELIABLE AND VALID MEASURES OF THREAT DETECTION PERFORMANCE IN X-RAY SCREENING

8.1 ABSTRACT

Over the last decades, airport security technology has evolved remarkably. This is especially evident when state-of-the-art detection systems are concerned. However, such systems are only as effective as the personnel who operate them. Reliable and valid measures of screener detection performance are important for risk analysis, screener certification and competency assessment, as well as for measuring quality performance and effectiveness of training systems. In many of these applications the hit rate is used in order to measure detection performance. However, measures based on signal detection theory have gained popularity in recent years, for example in the analysis of data from threat image projection (TIP) or computer based training (CBT) systems.

In this study, computer-based tests were used to measure detection performance for improvised explosive devices (IEDs). These tests were conducted before and after training with an individually adaptive CBT system. The following measures were calculated: pHit, d', Δm, Az, A', $p(c)_{max}$. All measures correlated well, but ROC curve analysis suggests that "nonparametric" measures are more valid to measure detection performance for IEDs. More specifically, we found systematic deviations in the ROC curves that are consistent with two-state low threshold theory of Luce (1963). These results have to be further studied and the question rises if similar results could be obtained for other X-ray screening data. In any case, it is recommended to use A' in addition to d' in practical applications such as certification, threat image projection and CBT rather than the hit rate alone.

8.2 INTRODUCTION

Technological progress enabled state-of-the-art X-ray screening to become quite sophisticated. Current systems provide high image resolution and several image "enhancement" features (e.g. zoom, filter functions such as negative image, edge detection etc.). But technology is only as effective as the humans that operate it. This has been realized more and more in recent years and the relevance of human factors research has increased substantially. Note that during rush hour, aviation security screeners often have only a few seconds to inspect X-ray images of passenger bags and to judge whether a bag contains a forbidden object (NOT OK) or whether it is OK. Threat object recognition does largely depend on perceptual experience and training (Schwaninger & Hofer, 2004). Most object recognition models agree with the

view that the recognition process involves matching an internal representation of the stimulus to a stored representation in visual memory (for an overview see Graf, Schwaninger, Wallraven, & Bülthoff, 2002; Schwaninger, 2004a). If a certain type of forbidden item has never been seen before, there exists no representation in visual memory and the object becomes very difficult to recognize if it is not similar to stored views of another object. Besides the aspect of memory representation, image-based factors can also affect recognition substantially (for a detailed discussion see, Schwaninger, Hardmeier, & Hofer, 2004). When objects are rotated, they become more difficult to recognize (effect of view). In addition, objects can be superimposed by other objects, which can impair detection performance (effect of superposition). Moreover, the number and type of other objects in the bag challenge visual processing capacity, which can affect detection performance as well (effect of bag complexity).

While CBT can increase detection performance substantially (Schwaninger & Hofer, 2004), recent results suggest that training effects are smaller for increasing the ability to cope with image-based factors such as effects of view, superposition and bag complexity (Schwaninger, Hardmeier, & Hofer, 2004). However, this conclusion relies on the availability of reliable and valid measures of detection performance. For example the hit rate is not a valid measure for estimating the detection performance in a computer-based test in which screeners are exposed to X-ray images of passenger bags and have to take OK / NOT OK decisions. The reason is simple: A candidate could achieve a high hit rate by simply judging most bags as NOT OK. In order to distinguish between a liberal response bias and true detection ability, the false alarm rate needs to be taken into account as well. This is certainly part of the reason why signal detection theory (SDT) has been used for analyzing X-ray screening data (see for example McCarley, Kramer, Wickens, Vidoni, & Boot, 2004; Schwaninger & Hofer, 2004). In general, reliable and valid measures of detection performance are certainly very important for risk analysis, screener certification and competency assessment, as well as for measuring quality performance and effectiveness of training systems.

The objective of this study was to compare different measures of screener detection performance. As clearly shown by Schwaninger and Hofer (2004), detection performance depends substantially on perceptual experience and training – at least for certain types of threat items. In order to evaluate different performance measures we used computer-based tests before and after CBT. Using a baseline test and a test after the training period makes it possible to compare different detection measures with regard to their reliability and validity while taking effects of training into account. The

following detection measures were compared in this study: pHit, d', Δm, Az, A', $p(c)_{max}$. These measures and the corresponding detection models are summarized in the following section.

8.3 DETECTION MODELS AND PERFORMANCE MEASURES

Signals are always detected against a background of activity, also called noise. Thus, detecting threat objects in passenger bags could be described as typical detection task, where the signal is the threat object and the bag containing different harmless objects constitutes the noise. A correctly identified threat object corresponds to a hit, whereas a bag, which contains no threat item judged as being harmless, represents a correct rejection. Judging a harmless bag as being dangerous is a false alarm, whereas missing a forbidden object in a bag represents a miss. In every detection situation, the observer must first make an observation and then make a decision about this observation.

Signal detection theory (Green & Swets, 1966; MacMillan & Creelman, 1991) and threshold theories (Krantz, 1969; Luce, 1963) are two fundamentally different approaches of conceptualizing human perception. The main difference between the two approaches is that threshold theories suppose a theoretical threshold, whereas in SDT the concept of a threshold is rejected in favor of an adjustable decision criterion. Threshold theories can be coarsely divided into high (e.g. single high-threshold theory or double high-threshold theory) and low threshold theories. These approaches assert that the decision space is characterized by a few discrete states, rather than the continuous dimensions of SDT. Single high threshold theory predicts ROC curves which are often not consistent with experimental data (Gescheider, 1998; Snodgrass & Corwin, 1988). The other types of threshold theories are low threshold theories originally described by Swets (1961) and Swets, Tanner, and Birdsall (1955). The two-state low threshold theory is a slightly newer version by Luce (1963). This theory is often as consistent with the data as are SDT models. But because no single sensitivity measure exists for this theory, it is not widely applied (MacMillan & Creelman, 1991).

According to SDT, the subject's decision is guided by information derived from the stimulus and the relative placement of a response or decision criterion. An anxious person would set the criterion very low, so that a very small observation would lead to a signal response. In contrast, another person might set the criterion very high, so that the sensory observation needs to be very strong that this person would give a "signal present answer". It is important to note that different persons can have different criterion locations, and it is also possible that the same person changes the location of

the criterion over time. For example the day after 9/11, many airport security screeners have moved their criterion in the direction that at the smallest uncertainty, they judged passenger bags as being dangerous. Although the hit rate increases, detection performance stays more or less stable, because also the false alarm rate increases.

In contrast to signal detection theory, threshold theories suppose that not the locus of the criterion causes the answer but a theoretical threshold. In the two-state low threshold theory by Luce (1963), the threshold is assumed to exist somewhere between the middle and the upper end of the noise distribution. During a sensory observation, an observer is in the detect state if the observation exceeds threshold and in the nondetect state if the observation is below threshold. The response one makes in either state may be biased by nonsensory factors. A person can say yes when in the nondetect state or say no when in the detect state. Manipulating variables such as payoff and signal probability changes the observers' response bias when they are in either one of the two possible detection states. The main disadvantage of this low threshold theory is its lack of a single sensitivity measure that can be calculated from hits and false alarms.

Different signal detection measures and sensitivity measures coming from threshold theories exist. One of the most popular and very often used parametric SDT measures is d'. It is calculated by subtracting the standardized false alarm rate from the standardized hit rate. A detection performance of d' = 0 means that the screener had exactly the same hit and false alarm rate – in other words that this screener just guessed. This measure may only be calculated under the assumption that the theoretical signal-plus-noise distribution and the noise distribution are 1) normally distributed (binormal ROC curves) and 2) that their variances are equal. These assumptions can be tested with receiver-operating characteristic (ROC) curves, where the proportion of hits is plotted as a function of the proportion of false alarms at different locations of the criterion. Maximum likelihood (ML) estimation algorithms for fitting binormal ROC curves are available (see Dorfman & Alf, 1969; Metz, 1989; Swets & Pickett, 1982). The second assumption can be tested with the slope of the standardized ROC curve (if the variances are equal the slope of the standardized ROC curve is 1). If the variances are unequal, another signal detection measure, Δm, is often used. One disadvantage of this measure is that it can only be computed when ROC curves are available. D' and Δm express sensitivity in terms of the difference between the means of the noise and signal-plus-noise distribution expressed in units of the noise distribution. If the ROC curves are not binormal, it is still possible to express sensitivity as the area under the ROC curve.

Another well known measure, which is "nonparametric" (or sometimes also called "distribution-free") is A' and was first proposed by Pollack & Norman (1964). The term "nonparametric" refers to the fact that the computation of A' requires no a priori assumption about underlying distributions. A' can be calculated when ROC curves are not available and the validity of the normal distribution and equal variance assumptions of the signal-noise and noise distribution can not be verified.

A' can be calculated by the following formula (Grier, 1971): $A' = 0.5 + [(H - F)*(1 + H - F)]/[4H*(1 - F)]$, whereas H is the hit rate and F the false alarm rate. If the false alarm rate is greater than the hit rate the equation must be modified (Aaronson & Watt, 1987; Snodgrass & Corwin, 1988): $A' = 0.5 - [(F - H)*(1 + F - H)]/[4F*(1 - H)]$.

As Pastore, Crawley, Berens, and Skelly (2003) have pointed out, this does not mean that these measures are an accurate reflection of their theoretical origin (i.e. that A' reflects the area under a reasonable ROC curve) or that A' is a distribution-free measure or fully independent of a response bias (see also MacMillan & Creelman, 1996). Thus, the term "nonparametric" is somewhat misleading. A further disadvantage of A' is that it underestimates detection ability by an amount that is a function of the magnitude of bias and decision ability (Pastore et al., 2003). But because A' can be easily computed and no further assumptions on the underlying noise and signal-plus-noise distribution have to be made, researchers often use this measure when the assumptions of SDT are not fulfilled or cannot be tested. Another measure, which is sometimes used, is the unbiased proportion correct $p(c)_{max}$. This measure can be calculated from d' and used instead of A' (see MacMillan & Creelman, 1991; Pastore et al., 2003). As d', it is independent of any response biases. Whenever using p(c)max, double high-threshold theory is implied (for detailed information on double-high threshold theory see Egan, 1958; summarized in Green & Swets, 1966).

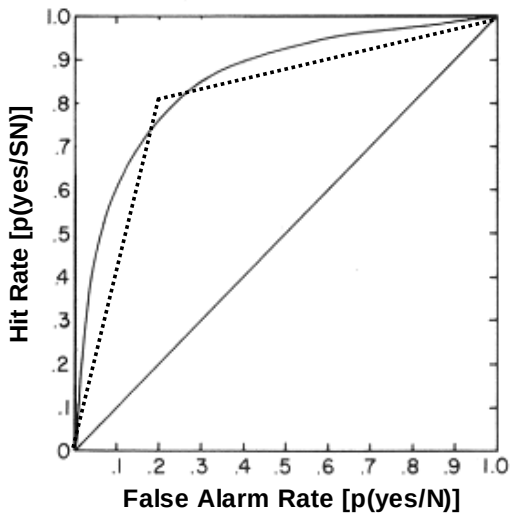

Figure 1. ROC implied by signal detection theory (solid curve) and by two-state low threshold theory from (Luce, 1963) (dashed lines).

To investigate which performance measures for threat detection in X-ray images are valid and reliable for novices as well as for experts it is important to examine the form of ROC-curves prior and after training. Note that signal detection and threshold theories predict different forms of ROC curves. In

linear coordinates the ROC curve predicted from the two-state low threshold theory of Luce (1963) are two straight lines, whereas the ROC curve predicted from signal detection theory is a symmetrical curve (see Figure 1).

8.4 METHOD

We used a computer-based training (CBT) system for improvised explosive devices (IEDs), which was developed based on object recognition theories and visual cognition. For detailed information on this CBT system (X-Ray Tutor) see Schwaninger & Hofer, 2004 and Schwaninger (2003, 2004b).

8.4.1 PARTICIPANTS

The original sample size of participants was seventy-two (fifty females) at the age of 23.9 – 63.3 years (M = 48.3 years, SD = 9.0 years). Data of ten participants were not included in the analyses of this study because at least for one test date the slope of the standardized ROC curve was between -0.1 and 0.1. Thus, for the analyses in this study data of sixty-two participants were used. None of the participants had received CBT before.

8.4.2 TRAINING DESIGN

The detailed design of the data collection, design and material can be found in an evaluation study of the CBT published recently Schwaninger & Hofer (2004). In summary, four groups of participants had access to the CBT from December 2002 to May 2003. There were four training blocks counterbalanced across four groups of trainees using a Latin Square design. Prior to each training block, performance tests were taken containing the IEDs of the following training block. This method allowed to measure training effectiveness for IEDs never seen before in a standardized way.

Training and testing blocks consisted of sixteen IEDs. For the training, sixteen difficulty levels were constructed by combining each IED with bags of different complexities. At test, only the two most difficult combinations of IEDs and bags were used. To test training effects for different display durations, each bag was presented for 4 and 8 seconds.

All four tests consisted of 128 trials: 16 (IEDs) * 2 (two most difficult levels) * 2 (4 & 8 sec display durations) * 2 (harmless vs. dangerous bags). The order of trial presentation was randomized. For each trial, participants had to judge whether the X-ray image of the bag contained an IED (NOT OK response) or not (OK response). In addition, confidence ratings from 0 (very easy) to 100 (very difficult) were assessed after each decision.

For the purposes of this study we analyzed data of the first detection performance test conducted in Dec/Jan 2003 (Test 1, prior training) and data

of the third detection performance test conducted in March/April 2003 (Test 3, after 20 training sessions on average)[1].

8.5 RESULTS

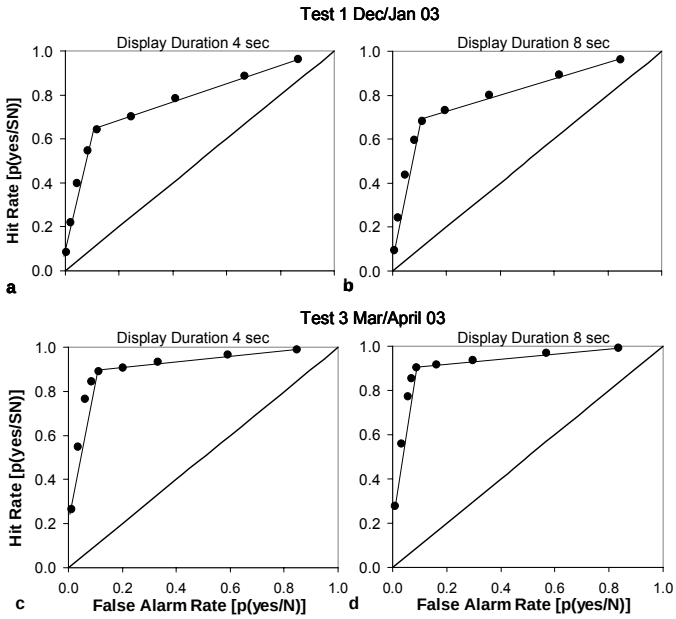

Figure 2. Unstandardized ROC curves for the two test dates (prior and after training), based on pooled data from 62 participants. a, b) ROC curves prior training with display durations of 4 sec (a) and 8 sec (b). c, d) ROC curves after training with display durations of 4 sec (c) and 8 sec (d).

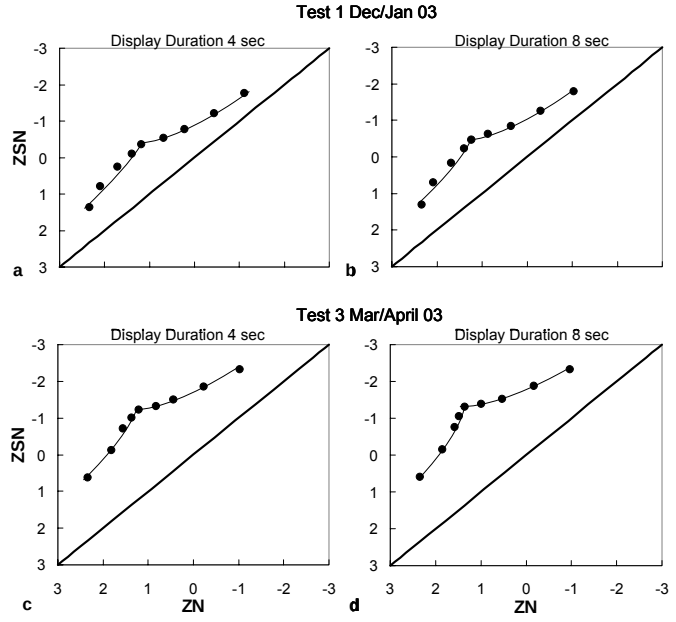

Figure 3. Standardized detection ROC curves for the two test dates (prior and after training), based on pooled data fort he 62 participants. a, b) ROC curves prior to training for 4 sec (a) and 8 sec (b). c, d) ROC curves after training for 4 sec (c) and 8 sec (d).

In order to plot ROC curves, confidence ratings were divided into 10 categories ranging from 1 (bag contains no IED for sure) to 10 (bag contains an IED for sure). Figure 2 shows the pooled unstandardized ROC curves prior training and after 20 training sessions for display durations of four (a, c) and eight seconds (b, d). As can be seen in Figure 2, the ROC curves seem to be better fitted by two straight lines than by a bimodal ROC curve as would be predicted from SDT.

As mentioned in section II such ROC curves are predicted from two-state low threshold theory (Luce, 1963) and are not consistent with the Gaussian distribution assumptions of SDT. Standardized ROC curves are shown in Figure 3 and one can clearly see that none of them is linear as would be predicted by SDT. This was confirmed in individual ROC analyses that revealed significant deviations from linearity for several participants ($\chi 2$-tests). Interestingly, nonlinearity seems to be more pronounced after training (see bottom halves of Figure 2 and 3).

These results suggest the

[1] Standard deviation was 8 training sessions.

existence of a low threshold and challenge the validity of SDT for explaining the data obtained in this study. As a consequence, the use of parametric SDT measures as reliable and valid estimates of threat detection performance might be questioned – at least as far as detection of IEDs in X-ray images of passenger bags is concerned.

However, it remains to be investigated whether our results can be replicated using other stimuli and whether similar results can be obtained when other threat categories are used (e.g. guns, knives, dangerous goods, etc.).

In any case it is an interesting question to what extent different detection measures correlate. Table 1 shows correlation coefficients (PEARSON) between the detection measures d', Δm, Az, A' and proportion correct $p(c)_{max}$ calculated according to MacMillan & Creelman (1991) for all four conditions (2 test dates and 2 display durations).

TABLE 1
CORRELATIONS BETWEEN DETECTION MEASURES

	Display Duration 4 sec (*r*)					Display Duration 8 sec (*r*)				
Test1 Dec/Jan 03	*d'*	pHit	Δm	Az	A'	*d'*	pHit	Δm	Az	A'
pHit	0.76					0.78				
Δm	0.89	0.74				0.89	0.77			
Az	0.84	0.74	0.95			0.81	0.78	0.93		
A'	0.75	0.73	0.75	0.77		0.80	0.83	0.81	0.88	
$p(c)_{max}$	0.97	0.74	0.88	0.87	0.75	0.94	0.82	0.57	0.82	0.81
Test3 Mar/April 03	*d'*	pHit	Δm	Az	A'	*d'*	pHit	Δm	Az	A'
pHit	0.74					0.74				
Δm	0.64	0.49				0.69	0.53			
Az	0.73	0.79	0.68			0.41	0.42	0.51		
A'	0.77	0.87	0.46	0.70		0.77	0.85	0.51	0.38	
$p(c)_{max}$	0.97	0.80	0.88	0.85	0.85	0.95	0.78	0.61	0.44	0.83

Note. All *p*-values < .01.

As can be seen in Table 1, the correlations between the different measures are quite high. There is a tendency of slightly smaller correlations after training.

Figure 4 visualizes the training effect using the different detection performance measures. A large training effect is clearly apparent for each detection measure. While substantial differences in slope can be observed between d' and Δm (Figure 4a), the comparison between A', Az and $p(c)_{max}$ reveals relatively parallel lines (Figure 4b). Detection performance is slightly better for display durations of 8 vs. 4 seconds, which is apparent for all measures.

Statistical analyses are reported only for A' since ROC analysis could not support parametric SDT measures. A two-way analysis of variance (ANOVA) with the two within-participants factors test date (prior and after training) and display duration (4 and 8 sec) showed a significant effect of test date, $F(1, 61) = 14.37$, $MSE = 0.001$, $p < .001$, and display duration, $F(1, 61) = 86.95$, $MSE = 0.53$, $p < .001$. Effect sizes (Cohen, 1988) were high, with $\eta^2 = .19$ for test date, and $\eta^2 = .59$ for display duration. There was no significant interaction between test date and display duration, $F(1, 61) = 3.24$, $MSE = 0.001$, $p = .08$.

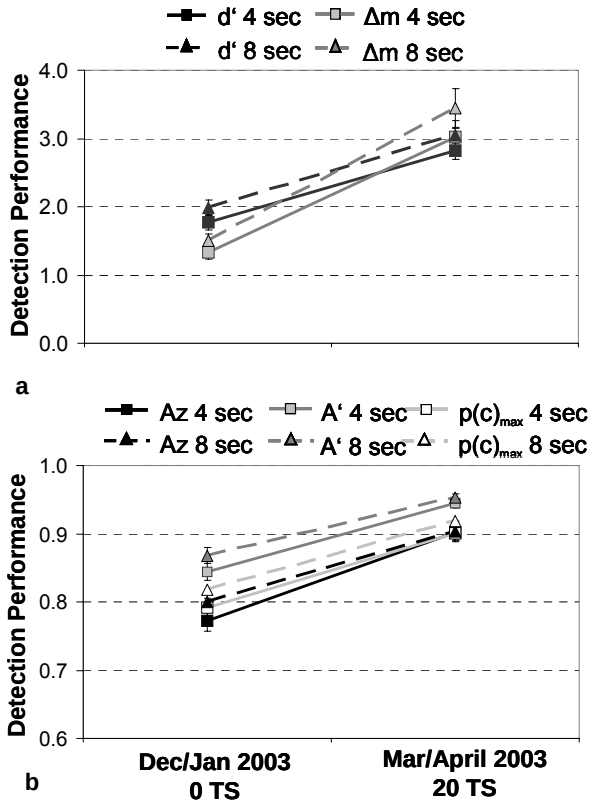

Figure 4. Illustration of training effect by comparing performance prior to training (Test 1, Dec/Jan 2003) and after 20 training sessions on average (Test 3, Mar/April 2003) for display durations of 4 and 8 sec. a) d', Δm, Az, A' and $p(c)_{max}$ TS = training sessions.

TABLE 2
RELIABILITY ANALYSES (CRONBACH'S ALPHA)

	Dec/Jan 2003 0 TS (Test 1)	Mar/April 2003 20 TS (Test 3)
Group A (N =12)	.92	.92
Group B (N = 15)	.90	.82
Group C (N = 17)	.90	.93
Group D (N = 18)	.91	.90

Note. Internal reliability coefficients broken up by participant group and test date.

Internal reliability was assessed by calculating Cronbach's Alpha using hits (NOT OK response for threat images) and correct rejections (OK responses for non-threat images). Table 2 contains the reliability coefficients for the two test dates (prior and after training) and each group of trainees while pooling display durations of 4 and 8 seconds.

8.6 DISCUSSION

The objective of this study was to compare different measures of X-ray detection performance while taking effects of training into account. To this end, computer based tests were used that were conducted before and after CBT.

From a regulators perspective the hit rate is sometimes the preferred measure when data from threat image projection (TIP) is used to judge the performance of screeners. However, the hit rate alone is not a valid measure

because it is not possible to distinguish between good detection ability and a liberal response bias. For example an anxious screener might achieve a high hit rate only because most bags are judged as being NOT OK. In this case security is achieved at the expense of efficiency, which would be reflected in long waiting lines at the checkpoint. It would be much more beneficial to achieve a high degree of security without sacrificing efficiency. This implies a high hit rate and a low false alarm rate. SDT provides several measures that take the hit and false alarm rate into account in order to achieve more valid measures of detection performance. Although the use of parametric SDT measures is still very common (e.g. Swets, 1996; McCarley et al., 2004; Schwaninger & Hofer, 2004), several studies have used "nonparametric" A' because its computation does not require a priori assumptions about the underlying distributions (e.g. Fisk & Schneider, 1981; Prkachin, 2003; Schwaninger et al., 2004). ROC analysis can be used to test whether the assumptions of SDT are fulfilled. We found that standardized ROCs deviated from linearity both before and after training of IED detection. Interestingly, unstandardized ROC curves could be fitted very well by two straight lines, just as would be predicted from two-state low threshold theory of Luce (1963). These results challenge the validity of SDT measures as estimates of threat detection performance, at least when detection of IEDs in X-ray images of passenger bags is concerned. It certainly remains to be investigated whether our results can be replicated with different stimulus material and other threat items than IEDs. In any case however, our findings suggest that other detection performance measures than those from SDT should be considered, too. As mentioned above, the calculation of A' requires no a priori assumption about the underlying distributions, which has often been regarded as an advantage over SDT measures such as d' and Δm. In many applications such as risk analysis, quality performance measurement, and competency assessment based on TIP or CBT data, only hit and false alarm rates are available and multipoint ROCs can not be obtained to test the assumptions of SDT measures. At least in these cases it should be considered to use A' in addition to d', while certainly both measures are more valid estimates of detection performance than the hit rate alone.

Finally, it should be noted that the five psychophysical measures compared in this study were usually strongly correlated. More specifically, the measures that are most often reported in the detection literature, A' and d', correlated in all four test conditions with r >= .75. And in a recent study using computer-based tests with different types of threat items even higher correlations between A' and d' were found ($r > .90$, Schwaninger et al., 2004).

8.7 REFERENCES

Aaronson, D., & Watt, B. (1987). Extensions of Grier's computational formulas for A' and B" to below-chance performance. *Psychological Bulletin, 102*, 439-442Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum, Hillsdale.

Dorfman, D.D., & Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals – rating method data. *Journal of Mathematical Psychology, 6*, 487-496.

Egan, J.P. (1958). Recognition memory and the operating characteristic, Hearing and Communication Laboratory, *Technical Note AFCRC-TN-58-51*, Indiana University.

Fisk, A. D., & Schneider, W. (1981). Control and Automatic Processing during Tasks Requiring Sustained Attention: A New Approach to Vigilance. *Human Factors, 23*, 737-750.

Gescheider, G. A. (1998). *Psychophysics: The Fundamentals* (3rd Ed), Mahwah, NJ: Lawrence Erlbaum Associates.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme*, Cognitive Vision Systems - CogVis (IST-2000-29375).

Green, D. M., & Swets, A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Grier, J. B. (1971). Nonparametric indexes for sensitivity and bias: Computing formulas. *Psychological Bulletin,* 75, 424-429.

Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

Pastore, R. E., Crawley, E. J., Berens, M.S., & Skelly, M.A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556-569.

Pollack, I., & Norman, D.A. (1964). A non-parametric analysis of recognition experiments, *Psychonomic Science, 1*, 125-126.

Krantz, D. H. (1969). Threshold theories of signal detection. *Psychological Review, 76*, 308-324.

Luce, R. D. (1963). A threshold theory for simple detection experiments, Psychological review, 70, 61-79.

MacMillan, N.A, & Creelman, C. D. (1996). Triangles in ROC space: History and theory of "nonparametric" measures of sensitivity and response bias. *Psychonomic Bulletin & Review, 3,* 164-170.

MacMillan, N. A., & and Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: University Press.

Metz, C. E. (1989). Some practical issues of experimental design and data analysis in radiological ROC studies. *Investigative Radiology, 24*, 234-245.

McCarley, J. S., Kramer, A., Wickens, C. D., Vidoni, E. D., & Boot, W. R. (2004). Visual Skills in Airport-Security Screening. *Psychological Science, 15,* 302-306.

Pastore, R. E., Crawley, E. J., Berens, M.S., & Skelly, M.A. (2003). "Nonparametric" A' and other modern misconceptions about signal detection theory. *Psychonomic Bulletin & Review, 10*, 556-569.Snodgrass, J.G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General, 117*, 34-50.

Prkachin, G. C. (2003). The effects of orientation on detection and identification of facial expressions of emotion. *British Journal of Psychology, 94,* 45-62.

Schwaninger, A. (2003). Training of airport security screeners. *AIRPORT, 05*, 11-13.

Schwaninger, A. (2004a). Objekterkennung und Signaldetektion. In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 106-130). Bern: Huber.

Schwaninger, A. (2004b). Computer based training: a powerful tool to the enhancement of human factors, *Aviation security international*, February, 31-36.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex : WIT Press.

Swets, J. A. (1961). Is there a sensory threshold? *Science, 134,* 168-177.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics – Collected Papers,* Mahwah, NJ: Lawrence Erlbaum Associates.

Swets, J. A., & Pickett, R.M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* New York: Academic Press.

Swets, J. A., Tanner, W. P. Jr., & Birdsall, T. G. (1955). *The evidence for a decision-making theory of visual detection*, Electronic Defense Group, University of Michigan, Tech. Rep. No. 40.

9 USING THREAT IMAGE PROJECTION DATA FOR ASSESSING INDIVIDUAL SCREENER PERFORMANCE

9.1 ABSTRACT

Threat image projection (TIP) is a technology of current x-ray machines that allows exposing screeners to artificial but realistic x-ray images during the routine baggage x-ray screening operation. If a screener does not detect a TIP within a specified amount of time, a feedback message appears indicating that a projected image was missed. Feedback messages are also shown when a TIP image is detected or in the case of a non-TIP alarm, i.e. when the screener indicated that there was threat but in fact no TIP was shown. TIP data is an interesting source for quality control, risk analysis and assessment of individual screener performance. In two studies we examined the conditions for using TIP data for the latter purpose. Our results strongly suggest using aggregated data in order to have a large enough data sample as the basis for statistical analysis. Second, an appropriate TIP library containing a large number of threat items, which are representative for the prohibited items to be detected is recommended. Furthermore, consideration should be given to image-based factors such as general threat item difficulty, viewpoint difficulty, superposition and bag complexity. Different methods to cope with these issues are discussed in order to achieve reliable, valid and standardized measurements of individual screener performance using TIP.

9.2 INTRODUCTION

The task of an airport security screener is to visually inspect passenger bags for forbidden or dangerous objects. In order to perform this task effectively, a screener needs to know which items are prohibited and what they look like in x-ray images of passenger bags. As pointed out by Schwaninger (2004a), some threat objects look very different in an x-ray image than in reality. Other prohibited items are difficult to identify in an x-ray image because they look similar to harmless objects. Because of these and other reasons, training and visual experience are essential in order to achieve and maintain a high level of detection performance (Schwaninger, 2003a; Schwaninger, 2004b; Schwaninger & Hofer, 2004).

In addition to such knowledge-based factors of expertise and training, there are several image-based factors, which also influence detection performance (Schwaninger, 2003b; Schwaninger, Hardmeier, & Hofer, 2004). When prohibited items are rotated they can become more difficult to recognize (effect of viewpoint). Superimposition by other objects in the bag can also affect detection performance (effect of superposition). In addition,

the number and type of other objects can affect the visual search for prohibited items (effect of bag complexity). Interestingly, comparable effects of these image-based factors on detection performance have been found for novices as well as for experts. Moreover, large inter-individual differences were found in the ability to cope with these image-based factors, which accounted for novices as well as for experts. Thus, these image-based factors seem to be rather related to relatively stable visual abilities than to training and visual experience.

The fact that image- and knowledge-based factors strongly influence detection performance points out that the effectiveness of aviation security technology is limited by the abilities and expertise of the humans that operate it. Therefore, reliable and valid procedures for assessing individual detection performance of screeners are relevant for quality control, risk analysis and screener certification purposes.

Large technological progress has been made in aviation security in the last two decades. One relatively new technology is threat image projection (TIP). This is a software function of state-of-the art x-ray machines that allows measuring of detection performance on the job. In TIP, virtual threat images are projected randomly on x-ray screening systems. For cabin baggage screening (CBS), fictional threat items (FTIs) are projected into x-ray images of real passenger bags in a random position. In hold baggage screening (HBS), combined threat images (CTIs) are displayed, i.e. virtual x-ray images of whole bags that can contain threat items. The use of CTIs is not possible in cabin baggage screening, because x-ray operators see the passengers and their luggage. Since in many hold baggage screening systems the operators are isolated from the passengers, it is possible to use CTIs in HBS TIP.

If a screener detects the projected threat item within a predefined time, the answer counts as hit. Missing a TIP-image is considered as miss. Non-Tip alarms are registered in CBS if a screener gives a threat present response when no TIP image was shown. In some HBS systems not only threat x-ray images but also non-threat x-ray images of passenger bags are shown. In this case, false alarms as well as correctly judging bags to be harmless are also written into TIP report files. Feedback messages are always presented to the screener when a TIP-image has been shown or in the case of a non-TIP alarm (CBS) or a false alarm (HBS).

TIP data is an interesting source for quality control, risk analysis and assessment of individual screener performance. Especially for the latter purpose, reliability of measurement is of special importance. This was examined in two studies using CBS and HBS data, respectively.

9.3 CBS STUDY

9.3.1 METHOD

9.3.1.1 TIP LIBRARY

A standard TIP library based on FAA (1997) was used, which is available on current TIP systems. A TIP:bag ratio of 1:50 was used in this study.

9.3.1.2 PARTICIPANTS

333 CBS airport security screeners took part in this study. They were all familiarized with the same TIP library using generic logins for several weeks before the study was started using individual logins. The study was conducted over a period of 7 months.

9.3.1.3 ANALYSES

There are different ways to estimate reliability. The most common procedures are test-retest, split-half, alternate forms and internal consistency analyses (for an overview see for example Kline, 2000; Murphy & Davidshofer, 2001). Because in both CBS and HBS current TIP software selects the threat items on a purely random basis, there can be quite substantial differences in repeated exposure to different items between screeners. It is therefore not possible to run the same TIP projections for every screener. This complicates reliability analyses because it implies that neither of the common reliability procedures can be applied in its pure form.

For the purpose of this study, two ways of data splitting were conducted: First, the hit rate of even days was correlated with the hit rate of odd days. Aggregated data was used, which was collected over a period of seven months. Second, the hit rate from one, two and three successive months was correlated with the hit rate of the following one, two and three months, respectively. For both ways of data splitting, some items can be in both halves, whereas some other items are only found in one of the two halves (varying across participants). Therefore, the reliability analyses in this study are a combination of split-half and test-retest reliability.

Psychophysical measures such as d' or A' are more valid estimates of detection performance than the hit rate alone (Hofer & Schwaninger, 2004; MacMillan & Creelman, 1991; Green & Swets, 1966). These measures take the hit *and* false alarm rate into account. In this study, all reliability analyses were done only with the hit rate due to the following reasons: First, it is not possible to get a valid false alarm rate from CBS TIP reports because the individual non-TIP alarm rate does not completely match the individual false alarm rate. If a screener detects a real threat in a bag when no TIP image is present, this is recorded as a non-TIP alarm in the TIP report. In this case the

response should count as a (true) hit and certainly not as a false alarm. Second, because correctly judging a bag to be harmless is not written into the CBS TIP report, the individual non-TIP alarm rate has to be estimated based on the averaged TIP to bag ratio, which can further reduce the internal validity of the estimates. Therefore, only the hit rate was analyzed as a measure of performance. Non-TIP alarm rates (CBS) and false alarm rates (HBS) are reported here to illustrate differences between individuals in their response tendencies.

9.3.2 RESULTS

9.3.2.1 CORRELATIONS BETWEEN HIT RATES OF EVEN AND ODD DAYS

Figure 1 shows the correlations between the hit rates of even and odd days, aggregated over different numbers of months and averaged between the categories guns, knives and IEDs[1].

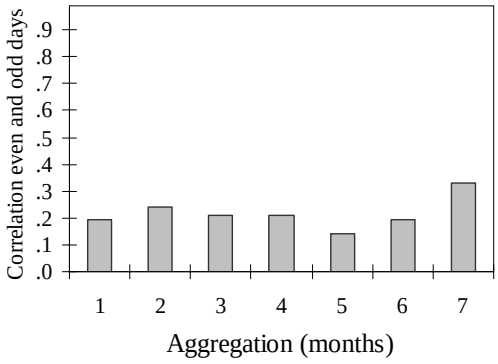

Figure 1. Correlations between even and odd days, aggregated over different numbers of months.

As can be seen in Figure 1, the correlations between the hit rate of even and odd days was relatively small and clearly below .40, even if TIP data from seven months were used. The mean even-odd day correlation for data of one month was $r = .19$, over seven months $r = .33$. It is important to note that the hit rate was very high and also very stable for all numbers of months (not shown in Figure 1). Moreover, the standard deviations were very small. These results suggest that most screeners achieved ceiling performance already in the first weeks when generic logins were used prior to the data collection for this study (see also section 9.3.1.2).

9.3.2.2 CORRELATIONS BETWEEN THE HIT RATE OF CONSECUTIVE MONTHS

Figure 2 illustrates the correlations for data of two, four and six successive months, calculated by correlating the hit rate between the first and second month, the hit rate between the first two and second two months, and the hit rate between the first three and second three months.

Correlation varied between $r = .32$ and $r = .58$. Thus, splitting the CBS data between different months resulted in higher correlations than splitting the data into even and odd days.

[1] Separate results for each category (guns, knives and IEDs) are very similar to the overall result and therefore not reported here.

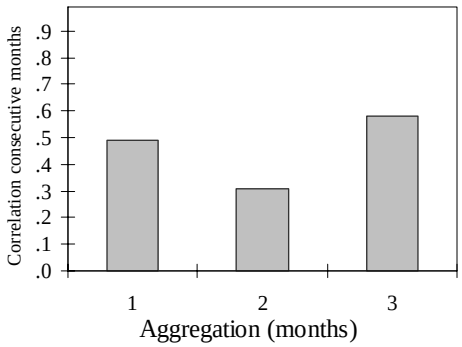

Figure 2. Correlations between the hit rate of the first and second month (1), of the first two and second two months (2), and of the first three and second three months (3).

9.3.2.3 INDIVIDUAL NON-TIP ALARM RATES

As can be seen in Figure 3, non-TIP alarms varied substantially between individual screeners. We found 0 percent for the screener with the lowest and 19percent for the screener with the highest non-TIP alarm rate.

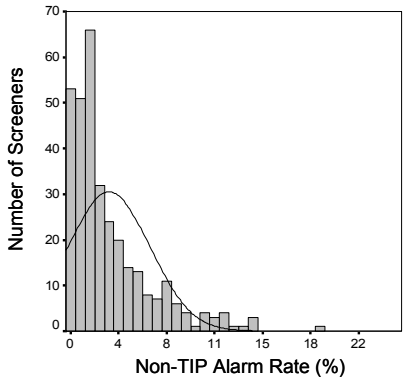

Figure 3. Distribution of individual non-TIP alarm rates (averaged over the 7 months period for each screener).

9.3.3 DISCUSSION

The correlations for the CBS TIP data of the standard TIP library available on conventional x-ray screening systems were clearly below .4 when splitting the data into even and odd days. When splitting the data between consecutive months, the correlations were higher, but still too small to conclude that this data is reliable enough for individual performance assessment (all $r <= .58$). Note however, that the hit rate was very high and stable over different months. In addition, very small standard deviations were observed. Thus, a ceiling effect in the data and small inter-individual differences could have resulted in small reliability coefficients.

The non-TIP alarm rate varied substantially between individual screeners, which reflects differences in response bias. Since the hit rate is dependent on individual response biases, its validity for measuring detection performance in terms of sensitivity is reduced.

9.4 HBS STUDY

9.4.1 METHOD

9.4.1.1 TIP LIBRARY

The library used for HBS TIP consisted of 1028 combined threat images (CTIs). 64 improvised explosive devices (IEDs) were selected by police experts from a large x-ray image database in order to create a representative sample of different IED types. Each IED was combined with 8 bags of different difficulties rated by 8 x-ray screening experts. Each bag was also displayed without the IED. Thus, the whole HBS TIP library consisted of 64 IEDs * 8 difficulty levels * 2 (harmless vs. dangerous bags) = 1028 CTIs. A TIP:bag ratio of 1:30 was used in this study.

9.4.1.2 PARTICIPANTS

74 HBS airport security screeners participated in this study. They were all familiarized with TIP using individual logins several weeks before the individual measurement started. The TIP images used in the introductory phase were different from the ones used for the reliability analyses. The study was conducted over a period of 16 months.

9.4.1.3 ANALYSES

The same method as for CBS was used to assess the reliability of HBS TIP data. Again, correlations between the hit rate of even and odd days for different numbers of months were calculated and correlations between the hit rates of several successive months were computed.

9.4.2 RESULTS

9.4.2.1 CORRELATIONS BETWEEN DATA OF EVEN AND ODD DAYS

Figure 4 shows the mean correlations between even and odd days aggregated over one month (r = .70) up to 16 months (r = .94). The hit rates (not shown in Figure 4) were smaller than in study 1. Moreover, much larger standard deviations were observed now.

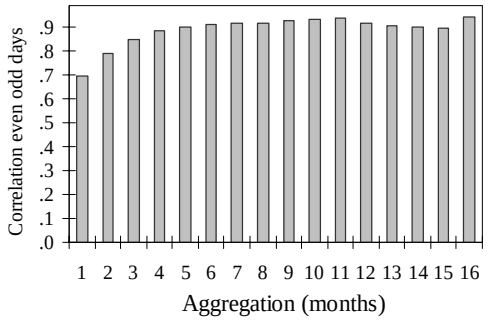

Figure 4. Correlations between even and odd days, aggregated over different numbers of months.

9.4.2.2 CORRELATIONS BETWEEN THE HIT RATE OF CONSECUTIVE MONTHS

Figure 5 shows the reliability coefficients calculated by correlating the hit rate between the first and second month, the first two and the following two months, the first three and the following three months etc.

The correlation between the hit rate of the first two months was $r = .60$, the correlation between the first eight and the following eight months was $r = .75$

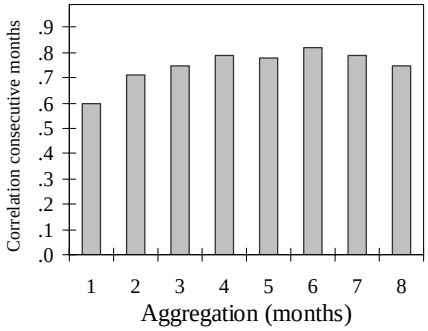

Figure 5. Correlations between the hit rates aggregated over different numbers of consecutive months.

9.4.2.3 INDIVIDUAL FALSE ALARM RATES

As explained in the method section, in this study half of all TIPs were harmless bags whereas the other CTIs contained an IED. Figure 6 shows the distribution of false alarm rates based on TIP trials containing a harmless bag (averaged for each screener over the 16 months period).

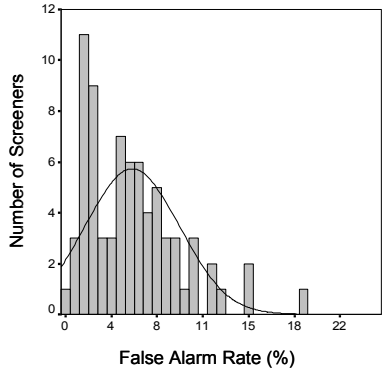

Figure 6. Distribution of individual false alarm rates from non-threat TIP trials (averaged across all 16 months for each participant).

9.4.3 DISCUSSION

Compared to the CBS TIP data, much higher correlations were revealed for the HBS TIP library used in this study. The correlation between even and odd days was .70 for data aggregated over one month and > .8 when data were aggregated over three months. For data aggregated over 6 or more months, the correlation between hit rates of even and odd days was > .9. When correlating the hit rate between different numbers of consecutive months, correlations were still quite high, although less high than when splitting the data into even and odd days. Again, as for the CBS data, screeners vary substantially in their response bias, which was reflected by the

variation in the false alarm rate of the HBS TIP data. Since the hit rate is affected by response bias, this result questions the validity of the hit rate for measuring detection performance in terms of sensitivity.

9.5 General Discussion

Threat image projection is a technology of current x-ray machines that allows exposing screeners to artificial but realistic x-ray images during the routine baggage x-ray screening operation. Because TIP allows realistic on the job measurement, it could be a useful tool for assessing individual screener x-ray detection performance. To this end, the measurement has to fulfill international standards of testing, i.e. the method used needs to be reliable, valid, objective and standardized. In this study, we analyzed TIP data from CBS and HBS in order to investigate reliability. Current TIP software allows only random projection of images. Since common reliability procedures need a standardized and controlled item set, they cannot be applied in their pure form. Therefore, we used a mixture between split-half and test-retest methods to estimate TIP data reliability. Data splitting was done in two different ways: The hit rate of even and odd days was correlated, while aggregating data over different numbers of months. Second, reliability was estimated by computing the correlation between the hit rate of consecutive months, e.g. the correlation between the first and second, the first two and second two, the first three and second three months, etc.

In the first study, the standard CBS TIP library was used, which is available on conventional x-ray screening systems. We found very low reliability values (all r <=.58), even for data aggregated over seven months. This is true for both data splitting methods used in this study. Although in general, splitting the data into different numbers of successive months resulted in slightly larger correlation coefficients as when splitting data into even and odd days. It is important to note that the hit rate was very high and only a small inter-individual variance was observed.

In the second study, a more difficult image library was used in HBS. The correlation between even and odd days is already .70 for data aggregated over one month. For data aggregated over 6 or more months, the correlation is > .9. When correlating the hit rate between different numbers of consecutive months, correlations are still quite high, although lower than when correlating even and odd days. Compared to the CBS data of study 1, the hit rate of the HBS TIP data was not at ceiling, and larger standard deviations could be observed.

What reasons could account for the fact that reliability of CBS data was so low while reliability coefficients were much higher for HBS data? One reason

for the low reliability of CBS data could be a ceiling effect and the small inter-individual differences. When using TIP for individual performance assessment a large image library containing a representative sample of items of varying difficulty should be used. At least from a testing psychology standpoint it could also be considered to eliminate the most easy and most difficult items. Another reason for the higher reliability of HBS TIP data could be related to the differences in the TIP system. In HBS, combined threat images are used, i.e. the threat item is shown always with the same bag, embedded at the same position. Therefore, any effects of superposition and bag complexity are kept constant. In CBS, only the threat items are projected into x-ray images of real passenger bags. This induces much additional variance of image difficulty because luggage varies in terms of bag complexity. Moreover, depending on the randomly selected location of the FTI, large differences in superposition are found. Therefore, it remains to be seen, whether reliable data can be obtained using CBS TIP. We are currently investigating this in a study using a larger CBS TIP library that contains more difficult threat items.

This study also showed that there are substantial inter-individual differences in non-TIP alarm rates (CBS) and false alarm rates (HBS). This indicates differences in response bias, which also affects the validity of hit rates as a measure of detection. It certainly would be desirable to design TIP systems in which a valid measure of false alarm rate can be obtained so that more valid detection measures such as d' and A' can be derived from hits and false alarms (Hofer & Schwaninger, 2004; MacMillan & Creelman, 1991; Green & Swets, 1966). This is already possible today in HBS TIP because CTIs are used and harmless bags can be projected in order to obtain valid false alarm estimates. In CBS this is not possible because FTIs are projected into x-ray images of real passenger bags. When a screener detects a real threat item, a non-TIP alarm is recorded, even though in this case it is a (true) hit and certainly not a false alarm. By separately recording these cases more valid hit and false alarm rates could be calculated. The true hits would simply be added to the hits obtained in TIP, the false alarm rate would equal the non-TIP alarms minus the true hits. Based on corrected hit and false alarm rates it would be possible to calculate d' or A' scores, which are more valid detection measures than the hit rate alone.

9.6 REFERENCES

Federal Aviation Administration, (1997). Functional requirements for threat image projection systems on X-ray machines, *DOT/FAA/AR-97/67, August*.

Green, D. M., & Swets, A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hofer, F. & Schwaninger, A. (2004). Reliable and valid measures of threat detection performance in X-ray screening. *IEEE ICCST Proceedings, 38,* 303-308.

Kline, P. (2000). *The Handbook of Psychological Testing* (2nd edition), London: Routledge, 2000.

MacMillan, N. A., & and Creelman, C. D. (1991). *Detection theory: A user's guide.* Cambridge: University Press.

Murphy, K. R., & Davidshofer, C. O. (2001). *Psychological testing: principles and applications.* New Jersey: Prentice-Hall.

Schwaninger, A. (2003a). Training of airport security screeners. *AIRPORT, 05,* 11-13, 2003.

Schwaninger, A. (2003b). Evaluation and selection of airport security screeners. *AIRPORT, 02,* 14-15.

Schwaninger, A. (2004a). Increasing effectiveness and efficiency in airport security screening, WIT Transactions on the Built Environment.

Schwaninger, A. (2004b). Computer based training: a powerful tool to the enhancement of human factors. *Aviation Security International, FEB/2004,* 31-36.

Schwaninger, A., Hardmeier, D., & Hofer, F. (2004). Measuring visual abilities and visual knowledge of aviation security screeners. *IEEE ICCST Proceedings, 38,* 258-264.

Schwaninger, A. & Hofer, F. (2004). Evaluation of CBT for increasing threat detection performance in X-ray screening. In: K. Morgan and M. J. Spector, *The Internet Society 2004, Advances in Learning, Commerce and Security* (pp. 147-156). Wessex : WIT Press.

10 EVALUATION OF CBT FOR INCREASING THREAT DETECTION PERFORMANCE IN X-RAY SCREENING

10.1 ABSTRACT

The relevance of aviation security has increased dramatically in recent years. Airport security technology has evolved remarkably over the last decade, which is especially evident for state-of-the-art X-ray screening systems. However, such systems will be only as effective as the people who operate them. Recognizing all kinds of prohibited items in X-ray images of passenger bags is a challenging object recognition task. In this article we present a method to measure screener detection performance based on signal detection theory. This method is applied to measure training effects resulting from individually adaptive computer based training (CBT). We have found large increases of detection performance and substantial reductions in response time suggesting that CBT is a very effective tool for increasing effectiveness and efficiency in aviation security screening.

10.2 INTRODUCTION

Working at an aviation security checkpoint is an important and demanding task. This is especially evident for the X-ray screener who has only a few seconds of inspection time to decide whether an X-ray image of a passenger bag is OK or needs to be manually searched (NOT OK). The X-ray screening task can be described as a signal detection situation in which prohibited items represent the signal and the remaining visual information in the X-ray image of the bag represents noise. Screener detection performance can be calculated using sensitivity measures from signal detection theory such as d', Δm or Az (Green & Swets, 1966, MacMillan & Creelman, 1991). These measures are based on hit rates and false alarm rates and are relatively independent of response biases. This is of special importance for measuring detection performance in X-ray screening tests. If the false alarm rate is not considered, it is not possible to distinguish between good detection performance and a "liberal" response bias (Schwaninger, 2004a). This can be illustrated by a simple example (Schwaninger, 2003a). Let us assume that two screeners A and B take a test in which 200 X-ray images of passenger bags are shown and half of them contain prohibited items. Both screeners detect threat items in 90 percent of the cases (hit rate). When the bag contains no prohibited items, screener A judges bags as being NOT OK in only 11 percent of the cases (false alarm rate). In contrast, screener B has a false alarm rate of 78 percent. Whereas screener A has a high detection performance, screener B achieves a high hit rate at the expense of efficiency,

which would result in substantially longer waiting lines at the checkpoint. This difference becomes apparent when detection performance is measured by d' = z(H) – z(FA), whereas H denotes the proportion of hits and FA the proportion of false alarms. In the formula z denotes the z-transformation, i.e. H and FA are converted into z-scores (standard-deviation units). In the example mentioned above screener A would have a detection performance of d' = z(0.90) – z(0.11) = 2.51 whereas screener B has a d' = z(0.90) – z(0.78) = 0.51. In other words, detection performance of screener A is almost 5 times higher!

If a CBT system is effective, it should be expected that detection performance d' increases as a result of training. Moreover, if threat items are seen repeatedly during training it could also be expected that they become better represented in visual memory, which could result in faster response times.

However, several methodological considerations need to be taken into account in order to achieve reliable measurements of CBT effectiveness in terms of d' increases and response time decreases. Schwaninger, 2003b identified three image-based factors that influence X-ray detection performance. Threat items can be more or less superimposed by other objects (effect of superposition). Second, the number and type of other objects in a bag challenge visual search and recognition processes such that threat items in more "complex" bags usually result in a lower detection probability (effect of bag complexity). Third, when objects are rotated away from the canonical view (Palmer, Rosch, & Chase, 1981) they usually become more difficult to recognize (effect of viewpoint). Since these effects have been shown to affect detection performance (Schwaninger, 2003b), image-based difficulty of X-ray images needs to be carefully controlled in a longitudinal study designed for evaluating CBT effectiveness. Moreover, display duration could be an important variable as well and should therefore be varied. Finally, only X-ray images of bags and threat items that have not been seen during training should be used in order to measure CBT effectiveness reliably.

These considerations were taken into account using a pilot study, a pre-selection test, and a Latin Square counterbalanced design with four tests of equal difficulty and four groups of screeners with comparable average threat detection ability.

10.3 METHOD

The CBT used in this study was X-Ray Tutor, an individually adaptive training system based on object recognition and visual cognition (for recent

reviews on these topics see Graf, Schwaninger, Wallraven & Bülthoff, 2002; Schwaninger, 2004a). The main aim of the system is training object recognition by increasing the number and strength of view-based representations in visual memory. X-Ray Tutor is driven by software algorithms that monitor student performance and adjust images presented to provide threat types and bag difficulty needed for the student to learn and progress based on performance deficiencies. For further information see Schwaninger, 2003c and Schwaninger, 2004b.

10.3.1 Pilot Study

Threat images were created by combining X-ray images of improvised explosive devices (IEDs) with X-ray images of passenger bags using a customized TRX algorithm. In the pilot study, 4000 X-ray images were used, i.e. 2000 harmless bag images and 2000 threat images (125 IEDs * 16 bags per IED). Image difficulty was rated by eight expert screeners of Zurich Airport using a slider control (rating scale 0-100). Inter-rater reliability was estimated by calculating Cronbach's Alpha among raters. Alpha for IEDs (averaged across the 16 X-ray images) was .96. Alpha for X-ray images (without averaging) was .82. Images were ordered by average rated difficulty so that 16 difficulty levels were obtained per IED.

10.3.2 Training Library

In the training system 64 of the 125 IEDs were used. Thus, the training library consisted of 1024 X-ray images containing a bag with an IED (64 IEDs * 16 bag difficulty levels) and 1024 harmless X-ray images showing the same bags without IED.

10.3.3 Participants

Seventy-two screeners (fifty female) at the age of 23.9 – 63.3 years (M = 48.3 years, SD = 9.0 years) took part in this study. None of them had received a special IED or computer based training before. These screeners were divided into four groups (group A: N = 17, group B: N = 18, group C: N = 18, group D: N = 19) as described in the next paragraph.

10.3.4 Grouping of Participants

Prior to training, a pre-selection test was used to distribute the screeners among four groups of equivalent detection performance. To this end, 16 IEDs rated in the pilot study were used, which were not contained in the training library. Each IED X-ray image was combined with a bag image of medium and high difficulty (difficulty level 9 and 15 estimated in the pilot study as described in section 10.3.1). The entire pre-selection test consisted of

64 trials: 16 IEDs * 2 difficulty levels * 2 trial types (threat images vs. harmless bags). The order of image presentation was counterbalanced across screeners. The task of the screeners was to decide whether the presented luggage contained an IED or not. After each answer, they rated the difficulty of each image on a slider from 0 (very easy) to 100 (very difficult). Statistical analyses showed that the standardized ROC curve is best described by a linear trend, $R^2 = .93$, $p < .001$. Thus, the parametric detection performance measures Δm and d' (Green & Swets, 1966, MacMillan & Creelman, 1991) could be calculated for each screener and four groups of comparable mean detection performance were created (Table 1)[1]. Three of the 72 screeners did not participate in this pre-selection test because they were not available during the period of testing which lasted 32 days (compare the number of screeners in Table 1 and section 10.3.3).

Groups of screeners	Δm	d'	r
Group A (N = 16)	1.58 (0.76)	1.75 (0.70)	.90
Group B (N = 17)	1.98 (2.14)	1.87 (1.02)	.89
Group C (N = 18)	1.63 (0.82)	2.07 (0.91)	.88
Group D (N = 18)	1.58 (0.73)	1.90 (0.88)	.84

Table 1: Mean Δm, d' and their correlation, listed separately for each group of screeners. Values in parentheses represent standard deviations. All correlations (r) are significant with p <.01.

A one factor ANOVA with group as between-subjects factor confirmed that the created groups were comparable in terms of their detection performance. There were no significant differences, neither for the Δm-values, $F(3, 65) = 0.40$, $p = .75$, nor for the d'-values, $F(3, 65) = 0.38$, $p = .77$. For both measures, no post hoc pairwise comparison between groups reached a statistic significant value (all p-values >.25).

10.3.5 TRAINING BLOCKS

The 64 IEDs used for training were distributed among four blocks of 16 IEDs so that all blocks were of comparable mean difficulty according to the difficulty ratings of the pilot study. One training block consisted of 512 images, i.e. 16 IEDs * 16 bags (difficulty levels) * 2 trial types (threat images vs. harmless bag images).

Standardized measures of difficulty ratings were subjected to one-way repeated measures ANOVA with training block as within-subjects factor. This analysis confirmed that the four training blocks were of equal difficulty. There was no effect of training block, $F(1.72, 12.04) = 0.47$; $MSE = 0.004$; p

[1] This data analysis was done prior to the study in chapter 8 (study 4). Because the linear trend explained the data very well, and no further examination of the data was done at the time of this study, only parametric measures were used in this study to evaluate the training. It has to be noted that in study 4, all performance measures correlated very high and that the training effect could be shown also with the nonparametric measure A' (see study 4, Figure 4).

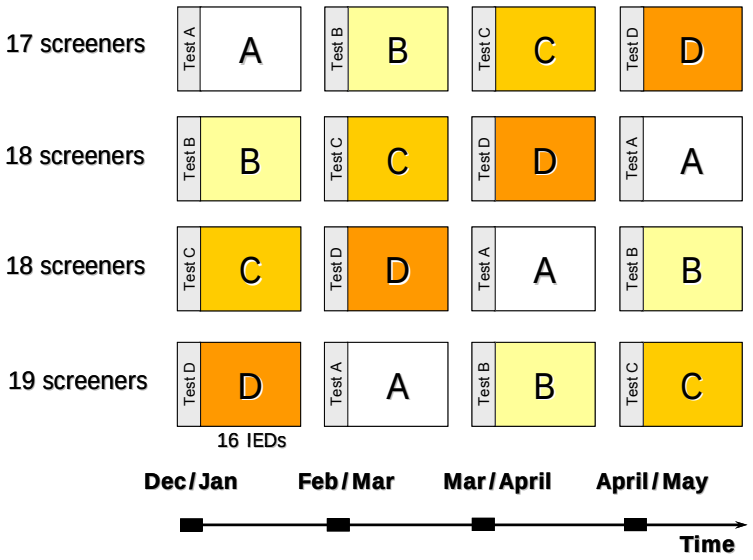

Figure 1: Latin Square design. A-D: Training blocks. Each training block consisted of 16 IEDs in 16 difficulty levels (bag images). Before each training block, a detection test containing the IEDs of the following training block was used to measure the training effects (see text for details). The study was carried out during six months starting December 2002 (see x-axis).

= .94 and pairwise comparisons between the training blocks showed no significant differences for any of the comparisons (all p-values > .25).

During training each IED was first presented in its easiest difficulty level. The order of IEDs was randomized across participants. The difficulty level was increased successively for each screener based on achievements in training (for more information on X-Ray Tutor see Schwaninger, 2003; 2004). Each training session was automatically terminated after 20 minutes.

The order of training blocks was counterbalanced across the four groups of trainees using a Latin Square design (see Figure 1). Between each training block the detection performance was measured in testing blocks. During training, each X-ray image was presented for a maximum of 8 seconds. Trainees had to decide whether the bag contains an IED or not by a clicking on one of two buttons. Subsequently, they judged the difficulty of the X-ray image from 0 (very easy) to 100 (very difficult) using a slider control. Screeners received immediate feedback to their answers. For X-ray images containing an IED the feedback messages were either "Threat detected" (hit) or "Threat missed" (miss). For innocent X-ray images the feedback messages were "False alarm" or "Bag OK" (correct identification of a harmless bag). In addition, an information window could be displayed which showed a labeled X-ray image and photograph of the IED.

10.3.6 Testing Blocks

The participants were always tested using IEDs they had never seen before. This was achieved using testing blocks which contained the IEDs from the next training block (see Figure 1). At test, each IED was presented for 4 and 8 seconds in bags of the two highest image difficulty levels (15 and 16). As in training, each bag was also presented without the IED in order to obtain a better signal detection measure.

As in training, participants judged whether the presented luggage is NOT OK (contained an IED) or OK (contained no IED) and subsequently rated the difficulty of each X-ray image using a slider control.

All four tests consisted of 128 trials: 16 IEDs * 2 display durations * 2 difficulty levels * 2 trial types (threat images vs. harmless bags). The order of presentation was randomized. In contrast to the training blocks, no feedback and no additional information about the IEDs was available during tests.

10.4 Results

10.4.1 Descriptive Statistics

There was a large increase in detection performance measured by signal detection d' (Figure 2a). In order to assess training effectiveness we calculated percent increase values as compared to baseline measurement (first test results), averaging the two display durations. Relative detection performance d' was increased by 70.76 percent (Figure 2b). This is a remarkable effect if it is taken into account that on average screeners took only 28 training sessions during the six months period (*SD* = 10 TS). Moreover, for a subgroup of 52 screeners, who on average took 31 training sessions (*SD* = 8 TS), the training effect was even more pronounced; relative detection performance was increased by 84.46 percent.

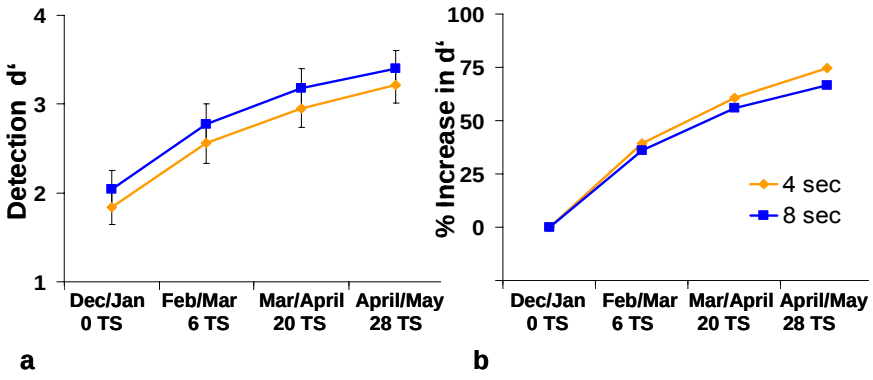

Figure 2: Absolute detection performance (a) and relative increase of detection performance (b) averaged across all 72 screeners. Display durations were 4 and 8 seconds. Error bars represent standard errors. TS = Number of training sessions.

10.4.2 Inferential Statistics

Only significant effects are reported using the conventional cut-off of $p < .05$. Effect sizes $\eta 2$ are reported and can be judged based on Cohen, 1988.

10.4.2.1 STATISTICAL ANALYSES OF DETECTION PERFORMANCE D'

Mean detection performance d' at the four test dates of each group are shown separately for the two display durations in Figure 3.

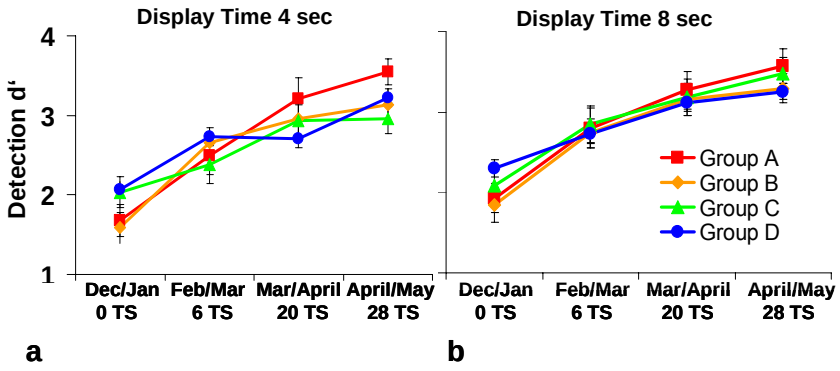

Figure 3: Detection performance d' of the four test dates for the four groups and the two display durations. Error bars represent standard errors. TS = Number of training sessions.

Again, the general training effect can be seen clearly. Detection performance d' of each group increased after each training block. A three-way analysis of variance (ANOVA) with the two within-subjects factors test date and display duration and the between-subjects factor group showed significant effects of test date, $F(2.81, 190.54) = 124.15$, $MSE = 0.44$, $p < .001$, and display duration, $F(1, 68) = 44.15$, $MSE = 0.14$, $p < .001$. With effect sizes of $\eta^2 = .65$ for test date and $\eta^2 = .39$ for display duration. The two-way interaction between test date and group was significant with an effect size of $\eta^2 = .10$, $F(9, 204) = 2.42$, $p < .05$. There was also a significant three-way interaction between test date, display duration and group, with an effect size of $\eta^2 = .09$, $F(9, 204) = 2.18$, $p < .05$.

In short, whereas the groups did not differ in their mean detection performance, there were slight differences in terms of how fast their detection performance increased across training when tested with 4 and 8 seconds of image presentation.

All Bonferroni-corrected pairwise comparisons between different test dates were significant confirming training effectiveness for the whole period of six months (all p-values $< .001$, with the exception of the comparison between test dates 3 (Mar/April) and 4 (April/May) with the p-value $< .01$).

10.4.2.2 STATISTICAL ANALYSES OF REACTION TIMES

Figure 4 (top) shows reaction times for bags containing an IED separately for the four screener groups and the two display durations of 4 and 8 sec.

Similarly, Figure 4 (bottom) depicts reaction times for harmless bags. Only reaction times of correct responses were analyzed. For threat images (bags with IED), a three-way ANOVA with test date and display duration as within-subjects factors and group as between-subjects variable showed a significant effect of test date, $F(1.71, 116.44) = 52.54$, $MSE = 1961145.60$, $p < .001$. The effect size was $\eta^2 = .44$. There was also a main effect of display duration ($\eta^2 = .57$), $F(1, 68) = 91.26$, $MSE = 476870.94$, $p < .001$. The two-way interaction between test date and display duration was also significant with $\eta^2 = .10$, $F(1.82, 123.72) = 7.30$, $MSE = 636920.93$, $p < .01$.

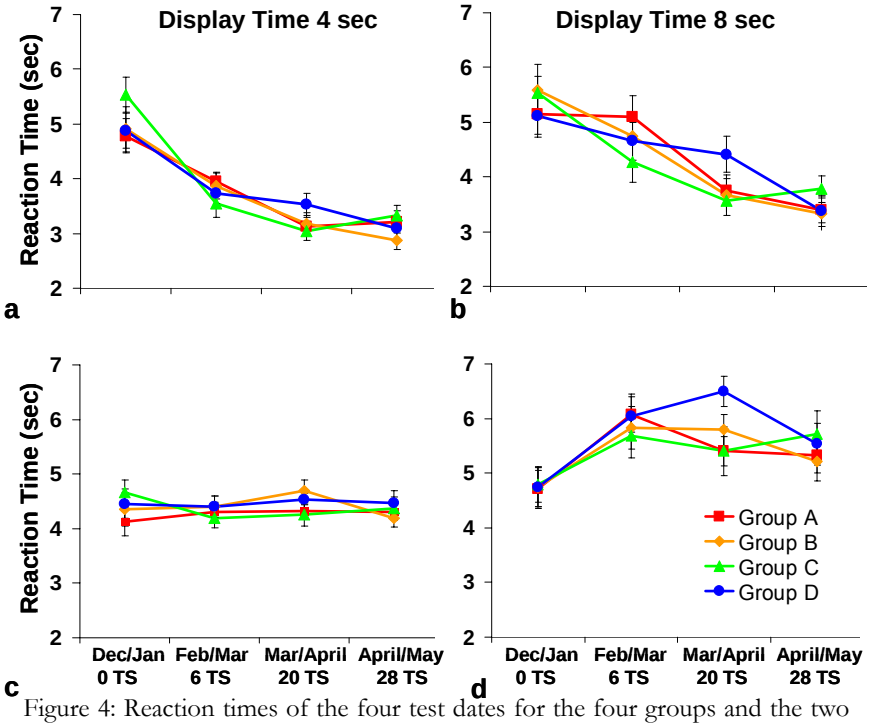

Figure 4: Reaction times of the four test dates for the four groups and the two display durations of 4 and 8 seconds. Top: Reaction times for bags containing an IED for 4 seconds (a) and 8 seconds (b). A clear decrease of reaction time was observed. Bottom: Reaction times for harmless bags for 4 seconds (c) and 8 seconds (d). Error bars represent standard errors. TS = Number of training sessions.

Bonferroni-corrected pairwise comparisons revealed significant differences for all comparisons between the reaction times of test dates (all p-values < .001, except for the comparison between test date 3 (Mar/April) and 4 (April/May) with $p < .05$). In short, response times for threat images decreased across training for all participant groups to a similar extend.

The same three-way ANOVA was used to analyze reaction times for harmless bags. Again, there was a main effect of test date, $F(1.99, 5.96) = 5.17$, $MSE = 2752802.22$, $p < .01$, with an effect size of $\eta^2 = .07$, which is much smaller than observed for threat images (see above). There was also a main effect of display duration ($\eta^2 = .74$), $F(1, 68) = 190.02$, $MSE = 901246.42$, $p < .001$, and a significant two-way interaction between test date and display duration ($\eta^2 = .28$), $F(2.49, 169.01) = 26.58$, $MSE = 463610.22.$, $p < .001$.

Except for the comparisons between test date 1 (Dec/Jan) and 2 (Feb/Mar) and 1 (Dec/Jan) and 3 (p-values < .05) no Bonferroni-corrected

pairwise comparison revealed significant differences between the reaction times of different test dates. Thus, in contrast to response times for threat images, there was no substantial reduction of response times for X-ray images of harmless bags.

10.5 DISCUSSION

The aim of this study was to develop a method in order to evaluate effectiveness of CBT for increasing threat detection performance in X-ray screening. Signal detection measures take the hit rate and the false alarm rate into account and provide more valid and reliable measures of detection performance than the hit rate alone (Schwaninger, 2003a, 2004a). ROC linearity analyses revealed that parametric measures d' and Δm can be computed (Green & Swets, 1966, MacMillan & Creelman, 1991). The two measures were strongly correlated as revealed in a pre-selection test that was used to create four groups of screeners with equivalent detection performance. Four tests of equal X-ray image difficulty were created based on difficulty ratings by eight expert screeners. Inter-rater reliability was sufficient suggesting that difficulty ratings could serve as estimates of objective detection performance.

A Latin Square counterbalanced design was used to measure CBT effectiveness in a longitudinal study of six months during which each screener took about 2 training sessions of 20 minutes per week. None of them had received a special IED or computer based training before. Only new X-ray images were used in the four tests in order to measure training effectiveness in terms of generalization to new threat items. Remarkable increases in detection performance d' were observed. Relative increase in detection performance d' as compared to the first test was 71 percent after an average of 28 training sessions during the six months period. For a subgroup of 52 screeners, who on average took 31 training sessions, relative increase in detection performance d' was even higher, i.e. 84 percent.

Image display duration at test had a small but reliable effect. When images were displayed for 4 seconds, performance was a bit worse than for 8 second display durations. This effect remained relatively stable across the four tests conducted during the six months period.

More interesting was the decrease in response time for detecting threat items as a result of training. This finding is consistent with the assumption that individually adaptive CBT increases the number and strength of view-based representations of threat items in visual memory and thus could explain a reduction of detection time. Since no response time reduction was observed for harmless bag images, the learning effect indeed seems to be

more related to visual memory representations than to increased general visual processing capacities.

In sum, the results of this study suggest that individually adaptive CBT is a powerful tool for increasing threat detection performance in X-ray screening of passenger bags.

10.6 REFERENCES

Green, D. M., & Swets, A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

MacMillan, N. A., & and Creelman, C. D. (1991). *Detection theory: A user's guide*. Cambridge: University Press.

Schwaninger, A. (2004a). Objekterkennung und Signaldetektion. In: B. Kersten & M.T. Groner (Eds.), *Praxisfelder der Wahrnehmungspsychologie* (pp. 106-130). Bern: Huber.

Schwaninger, A. (2003a). Evaluation and selection of airport security screeners. *AIRPORT*, 02, 14-15.

Schwaninger, A. (2003b). Reliable measurements of threat detection. *AIRPORT*, 01, 22-23.

Palmer, S.E., Rosch, E., & Chase, P. (1981). Canonical perspective and the perception of objects. In: J. Long & A. Baddeley (Eds.), *Attention and Performance IX* (pp. 135-151). Erlbaum: Hillsdale, N.J.

Graf, M., Schwaninger, A., Wallraven, C., & Bülthoff, H.H. (2002). Psychophysical results from experiments on recognition & categorisation. *Information Society Technologies (IST) programme*, Cognitive Vision Systems - CogVis (IST-2000-29375).

Schwaninger, A. (2003). Training of airport security screeners. *AIRPORT*, 05, 11-13.

Schwaninger, A. (2004). Computer based training: a powerful tool to the enhancement of human factors, *Aviation security international*, February, 31-36.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Erlbaum, Hillsdale.

11 DANKSAGUNG

Ich wurde während dieser Arbeit durch verschiedene Personen unterstützt und gefördert. Als erstes möchte ich mich bei meinem Betreuer Dr. A. Schwaninger bedanken, der mich nicht nur während dem Doktorat, sondern schon im Studium hervorragend betreute und in mir schon damals die Freude an der wissenschaftlichen Forschung geweckt hatte. Von ihm lernte ich, wie man wissenschaftliche Experimente plant und durchführt, und wie man danach diese Arbeiten so zu Papier bringt, dass daraus wissenschaftliche Publikationen entstehen. Vielen Dank, dass diese Arbeit in diesem Rahmen möglich wurde.

Prof. Dr. W. Marx möchte ich auch besonderen Dank aussprechen. Er hat mich schon als Studentin wertvoll unterstützt, und mich während des Doktorats als Erstreferent mit vielen wertvollen Kommentaren begleitet. Meinem zweiten Referenten Prof. Dr. F. Mast möchte ich auch ein Dankeschön sagen, nicht zuletzt weil er mich mit seiner Grundstudiumsvorlesung vor vielen Jahren für die Wahrnehmungspsychologie begeisterte.

Zudem möchte ich mich für die tolle Zusammenarbeit bei den beiden ersten Experimenten bei Stefan Ryf bedanken, der insbesondere bei methodischen Fragen eine grosse Hilfe war. Die Studie zum Top-Down Einfluss in Kapitel 5 wurde zusammen mit Stefan Michel durchgeführt. Bei Prof. Dr. H.H. Bülthoff bedanke ich mich ebenfalls für seine wertvollen Kommentare bei dieser Studie. Ebenfalls ein herzliches Dankeschön an Diana Hardmeier. Sie arbeitete bei den ersten beiden Studien des angewandten Teils dieser Dissertation (Kapitel 6 und 7) mit viel spannenden und wertvollen Diskussionen intensiv mit.

Viele dieser angewandten Studien hätten nicht ohne die tolle Zusammenarbeit mit der Kantonspolizei Zürich, Flughafenpolizei durchgeführt werden können. Auch den betreffenden Personen möchte ich Danke sagen.

Fürs Korrekturlesen dieser Arbeit bedanke ich mich zudem herzlich bei Denise Long.

Ebenfalls besonderen Dank geht an meine Eltern und meinen Bruder für die emotionale Unterstützung während der Doktoratszeit. Auch möchte ich die Gelegenheit nutzen, mich bei meinen Eltern für die finanzielle Unterstützung während des Studiums zu bedanken – dies hat mir sicherlich den Weg zur Dissertation eröffnet.

Zuletzt möchte ich allen Teilnehmern der Experimente und Studien ein herzliches Dankeschön mit auf den Weg geben.

12 CURRICULUM VITAE

WORKPLACE AND PERSONAL DATA

Name	Hofer Franziska
Workplace	University of Zurich Department of Psychology General Psychology (Cognition) Visual Cognition Research Group Klosbachstrasse 107 8032 Zurich Switzerland Phone +41 44 254 38 51 Email f.hofer@psychologie.unizh.ch
Nationality	Switzerland
Date and Place of birth	28.12.1975, Switzerland
Marital Status	Single

PROFESSIONAL EXPERIENCE

Since June 2004	Research Group Manager Visual Cognition Research Group, University of Zurich, Department of Psychology, General Psychology (Cognition)
Aug 2001 – June 2002	Research assistant at the University of Zurich, Department of Psychology, General Psychology (Cognition)
July – August 2001	Internship at the ETH Zurich, Institute of Hygiene and Applied Physiology, Department of Applied Vision Research
March – June 2000	Neuropsychological internship at the University hospital of Zurich, Institute of Neurology, Department of Neuropsychology
Oct 1999 – Feb 2001	Semester assistant at the University of Zurich, Institute of Psychology and Institute of Neuroinformatics
Feb 1999 – Dec 2001	Part-time job ‚Project Team Assistant', IBM Switzerland (20-40%), Zurich
Oct 1995 – Jan 1996	Internship at the Cantonal Hospital of St.Gallen, Department of internal medicine

EDUCATION

2002 – 2005	Doctoral student at the University of Zurich, Department of Psychology, General Psychology (Cognition), PhD Thesis: A Psychophysical Approach to Object Recognition and its Application in Airport Security
1996 – 2002	Studies of Psychology, Neurophysiology and Anthropology, University of Zurich Lizentiat: June 2002
1995	Center for English Studies (TOEFL), San Francisco, California
1990 – 1995	High school, Burggraben St.Gallen (Typus E)
1988 – 1990	Secondary school in Flawil (SG)
1982 – 1988	Primary school in Flawil (SG)

FURTHER EDUCATION

Sept 2002	Summer School on Advanced Methods in the Social Sciences: Structural Equation Modeling (SEM)
Jan 2006	University of Zurich: Introduction to MySQL and PHP